Subseasonal statistical forecasts of eastern U.S. hot temperature events

Vijverberg, Sem; Schmeits, Maurice; van der Wiel, Karin; Coumou, And D.I.M.

published in
Monthly Weather Review
2020

DOI (link to publisher)
10.1175/MWR-D-19-0409.1

document version
Publisher's PDF, also known as Version of record

Subseasonal Statistical Forecasts of Eastern U.S. Hot Temperature Events

SEM VIJVERBERG,[a] MAURICE SCHMEITS,[b] KARIN VAN DER WIEL,[b] AND DIM COUMOU[a]

[a] *Institute for Environmental Studies, VU Amsterdam, Amsterdam, Netherlands*
[b] *Royal Netherlands Meteorological Institute, De Bilt, Netherlands*

(Manuscript received 2 January 2020, in final form 10 September 2020)

ABSTRACT: Extreme summer temperatures can cause severe societal impacts. Early warnings can aid societal preparedness, but reliable forecasts for extreme temperatures at subseasonal-to-seasonal (S2S) time scales are still missing. Earlier work showed that specific sea surface temperature (SST) patterns over the northern Pacific Ocean are precursors of high temperature events in the eastern United States, which might provide skillful forecasts at long leads (~50 days). However, the verification was based on a single skill metric, and a probabilistic forecast was missing. Here, we introduce a novel algorithm that objectively extracts robust precursors from SST linked to a binary target variable. When applied to reanalysis (ERA-5) and climate model data (EC-Earth), we identify robust precursors with the clearest links over the North Pacific. Different precursors are tested as input for a statistical model to forecast high temperature events. Using multiple skill metrics for verification, we show that *daily* high temperature events have no predictive skill at long leads. By systematically testing the influence of temporal and spatial aggregation, we find that noise in the target time series is an important bottleneck for predicting extreme events on S2S time scales. We show that skill can be increased by a combination of 1) aggregating spatially and/or temporally, 2) lowering the threshold of the target events to increase the base rate, or 3) adding additional variables containing predictive information (soil moisture). Exploiting these skill-enhancing factors, we obtain forecast skill for moderate heat waves (i.e., 2 or more hot days closely clustered together in time) with up to 50 days of lead time.

KEYWORDS: Atmosphere–ocean interaction; Rossby waves; Stationary waves; Atmosphere–land interaction; Forecast verification/skill; Seasonal forecasting

1. Introduction

Subseasonal to seasonal (S2S) predictions offer society valuable information on weather-related risk, allowing decision-makers to initiate early warning action plans for extreme events (WMO 2017) and to optimize resource management (Vitart and Robertson 2018; Vitart et al. 2017). Predictability on these time scales stems from regularly varying climate phenomena or variables that are evolving at lower temporal frequencies compared to the regular, more chaotic, weather (Doblas-Reyes et al. 2013; Haustein et al. 2016; Krishnamurthy 2019). This predictability can be exploited by 1) initializing a dynamical model with these slowly evolving variables such as soil moisture, sea ice, snow cover and sea surface temperature (Jaiser et al. 2012; Seo et al. 2019; Vitart and Robertson 2018) or 2) select low-frequency variables directly as input for purely statistical forecasting models, using past climate data to train them (Kretschmer et al. 2017; Cohen et al. 2018; Totz et al. 2017; Nobre et al. 2019; Alfaro et al. 2006) 3) or a combination of both (Dobrynin et al. 2018).

S2S predictability can be improved by postprocessing the output of dynamical models, which is conventionally done by compensating for systematic biases (Finnis et al. 2012; Doblas-Reyes et al. 2013). Alternatively, statistical models can be directly trained to make S2S predictions and offer computational efficiency, flexibility, and the precursor time series can be further analyzed to provide process information (Runge et al. 2019). On S2S time scales, their forecast skill can be comparable to that of dynamical models (Hall et al. 2017). In this paper, we use the word *precursor* to refer to an anomalous pattern or geographical region, while the *precursor time series* refers to the time series that results from a dimensionality reduction of this pattern or region. A better understanding of important precursors can also help with the (bias) correction of dynamical models, either by using the precursors directly or by using the statistical model to subsample only the reliable forecasting pathways of the dynamical model output (Dobrynin et al. 2018; Strazzo et al. 2019).

The ocean is the most important source of long-term memory that interacts with the atmosphere (Frankignoul 1985; Kushnir et al. 2002; Kaspi and Schneider 2011; Putrasahan et al. 2013; Thomson and Vallis 2018). The atmospheric response to SST anomalies (SSTA) in the tropics is more direct and local (i.e., via thermally driven deep convection and associated latent heat release; Kushnir et al. 2002). In the midlatitudes, the lower specific humidity content and smaller Rossby radius of deformation hinders

Denotes content that is immediately available upon publication as open access.

Corresponding author: Sem Vijverberg, sem.vijverberg@vu.nl

DOI: 10.1175/MWR-D-19-0409.1

the formation of strong deep convection as seen in the tropics, resulting in a weaker direct and local atmospheric response to SSTA (Kushnir et al. 2002; Hewitt et al. 2017). The midlatitude atmospheric response to a SST anomaly is mainly driven by the adjustment to thermal wind balance and an indirect response due to eddy feedbacks (Nie et al. 2016). The latter makes the atmospheric response depend on the background zonal mean climate, and thus also on the season and location of the SSTA (Kushnir et al. 2002; Nie et al. 2016; Putrasahan et al. 2013).

These nuances for the atmospheric response, suggest that statistical forecasting tools should not solely rely on known modes of variability in the climate system, often referred to as climate indices (e.g., Hessl et al. 2004; Steptoe et al. 2018; Nobre et al. 2019). Those indices represent spatially large-scale, and temporally low-frequent processes. A priori, one does not know if they contain the information that is relevant for the target of interest. This reasoning is supported by statistical studies and dynamical model results (McKinnon et al. 2016; Deng et al. 2018).

Following this rationale, McKinnon et al. (2016) showed that ''hot day'' events in the eastern United States are preceded by a specific SST pattern over the Pacific Ocean. This SST pattern [called the Pacific extreme pattern (PEP)] was found by analyzing composite SST anomalies that co-occurred at certain lags with heat events. The PEP pattern is characterized by a zonally oriented tripole cold–warm–cold pattern in the Pacific at approximately 35°N, related to the forcing and/or amplification of a Rossby wave train (Wirth et al. 2018). The PEP was shown to outperform the conventional climate indices, such as the Pacific decadal oscillation (PDO) or the El Niño–Southern Oscillation (ENSO). Using the PEP pattern, the study claimed remarkable long-lead predictability up to 50 days lead time for extreme events defined at the daily time scale. In their main text, they assessed the skill of the PEP time series using only a single validation metric (area under the relative operating characteristic). However, it is generally recommended to use multiple skill metrics that measure different aspects of the forecast quality to assess predictability (Wilks 2011). Further, the study used a rectangular box over the tripole SST region to define the spatial extent of their precursor whereas a more objective SST pattern detection tool will verify if a link with the response variable is indeed robust and therefore might provide better physical understanding and more predictive skill (Bello et al. 2015; Kretschmer et al. 2017).

A response-guided statistical forecast tool that searches for precursors that explain the full variability of a continuous response (i.e., target) variable has been developed already (Kretschmer et al. 2017), but consequently, it cannot handle a binary target time series. Here, we introduce a novel response-guided algorithm that objectively extracts precursors that are directly related to our binary target variable (i.e., the eastern U.S. hot-day-event time series; section 2c). We train this algorithm on reanalysis ERA-5 and 160 years of data from the coupled ocean–atmosphere model EC-Earth.

In this paper, we present 1) a comparison between our response-guided algorithm and the PEP pattern of McKinnon et al. (2016), and their relation to the relevant climate indices (section 3b), and 2) the verification of hot-day forecasts, thereby stressing the importance of using multiple skill metrics (section 3c). Section 3d shows that forecast skill can be boosted by using temporal aggregation and lower-threshold events. To enable forecasts of events defined on a daily resolution, we no longer aggregate our target time series in time, but we use a window probability approach and we increase the signal-to-noise ratio by increasing the domain for spatial aggregation (section 3e). Although our focus lies on retrieving predictability from the ocean, in section 3f we also include additional information from soil moisture, since it is known to be a potentially important precursor of heat waves (Seneviratne et al. 2010; Miralles et al. 2014; Ardilouze et al. 2017).

2. Method

a. Data

Our analysis relies on data from the ERA-5 reanalysis, 1979–2018 [Copernicus Climate Change Service (C3S) 2017] and from the EC-Earth v2.3 earth system model (coupling between ocean, atmosphere, land surface and sea ice) (Hazeleger et al. 2012) with 160 yrs of simulated present-day climate (van der Wiel et al. 2019). We calculate the daily maximum 2-m air temperature (mx2t) in ERA-5 (0.25° × 0.25°) by calculating the daily maximum of the ''maximum 2-m temperature since previous post-processing,'' with a step size of 1 h. For SST in ERA-5 we use daily means on a 1° × 1° grid. We additionally use information from ERA-5 soil moisture (1° × 1°) for the forecasts [i.e., the volumetric soil water levels of the second (7–28 cm) and third (28–72 cm) layer of the land surface model]. To remove the seasonal cycle and the global warming trend (of which the strength might vary throughout the year), all variables are linearly detrended for each day-of-year. Because a single day-of-year across 40 years is insufficient to reliably estimate the climatological mean value and trend, we apply a 25-day rolling mean (using a Gaussian window with a standard deviation of 12.5) to the raw ERA-5 data. From the raw data, we then subtract climatological mean and trend based on the smoothed data.

For EC-Earth, we use daily mean T2m and SST data (1.125° × 1.125°). The coupled ocean–atmosphere climate model experiment consisted of 2000 years of simulated present-day weather, from this we sampled 160 years for our study. The selected years are not chronological, which is a desired property for making good splits between training and test data, because no interannual information is passed from the previous to the subsequent years. For more information on the model simulation setup, see van der Wiel et al. (2019). For EC-Earth, the seasonal cycle and a potential long-term trend is directly removed for each

day-of-year (no prior smoothing), since 160 years should be enough to reliably estimate the trend and climatological mean value.

b. *Defining the target variable and the Pacific extreme pattern*

We define our target variable following McKinnon et al. (2016) and determine it for ERA-5 and EC-Earth based on the detrended temperature data. The study period consists of the climatological 60 hottest days of year, ranging from 24 June to 22 August (McKinnon et al. 2016). The target variable is retrieved by, first, performing an objective identification of spatial clusters within the United States, where grid cells are clustered together if they tend to experience extreme events simultaneously. This clustering approach is expected to increase the signal-to-noise ratio and thereby helps to identify precursors; for more information on the clustering, see appendix A.

1) HOT-DAY EVENTS

McKinnon et al. (2016) calculated the spatial 95th percentile of daily maximum temperature anomalies within the eastern U.S. cluster. Hence, for each day, the spatial 95th percentile of all observations was calculated, which in practice means that each day contained the temperature value of only a single observation. This introduces some unwanted noise into the target time series since small-scale processes can affect the maximum temperature at a *single* observation and *single* moment in time. To improve the signal-to-noise ratio and at the same time stay close to the original definition, we calculate the spatial mean over the 10% warmest grid cells. This way we still end up with a very similar time series as compared with the T95 time series used by McKinnon et al. (2016) (Fig. 3, described in more detail below). We refer to this time series as $T90_m$ in the remainder of this article, with the lower case m referring to the spatial mean that is calculated. The hot-day time series (HD) is defined as

$$\mathrm{HD}(t) = 1 \quad \text{if} \quad T90_m > \left(\overline{T90_m} + \sigma_{T90_m}\right) \quad \text{else} \quad 0, \qquad (1)$$

with $\overline{T90_m}$ being the temporal mean and σ_{T90_m} being the standard deviation of $T90_m$. This results in a base rate of approximately 16%.

2) PEP

The PEP pattern is retrieved by taking the area weighted SSTA composite mean of hot-day events at lag τ. [The spatial region is defined by the rectangular box as depicted by green stippled lines in Fig. 4 (described in more detail below); the coordinates are 20°S–50°N and 145°E–130°W]. The PEP time series at lag τ is defined as the spatial covariance between the PEP pattern and the SSTA field at each time step [SSTA(t)]:

$$\mathrm{PEP}_\tau(t) = \frac{1}{N}\sum_i^N w_i\{[\mathrm{PEP}_\tau(t,i) - \overline{\mathrm{PEP}_\tau(t)}][\mathrm{SSTA}(t,i) - \overline{\mathrm{SSTA}(t)}]\}, \qquad (2)$$

where i denotes a grid cell of in total N grid cells within the rectangular box, w_i denotes the weight proportional to the gridcell area, and the overbar denotes the spatial mean.

c. *Composite-based precursor pattern algorithm*

This is a response-guided algorithm in the sense that it searches for a signal that directly relates to a response (i.e., target) variable of interest, in this case, the hot-day time series. It is inspired by the approach presented in McKinnon et al. (2016), who created a composite mean of hot-day events (i.e., calculating the mean SSTA that co-occurred with hot-day events at a certain lag). The null hypothesis would be that the SSTAs are unrelated to heat-wave events, meaning that one would be randomly sampling anomalies with respect to climatology, which should approximate zero. However, a distinct pattern of significantly deviating SSTA was found. This algorithm automatically infers the precursor regions based on robust anomaly patterns in the composite mean. The algorithm is described in step 1 and 2 (Fig. 1), and the parameters are listed in Table 1.

DETECTING ROBUST SST PRECURSORS

Robust anomalous grid cells should (i) be insensitive to the exclusion of a (number of) year(s) and (ii) SST anomalies should persist through time for at least a few days. Criterion (i) is tested by creating subsampled composite mean (SCM) maps and setting grid cells exceeding a percentile threshold (param = *SCM_perc_thres*; Table 1) to 1 and the rest to 0. We iteratively remove a number of training years based on a percentage. If we are, for example, removing 7.5% of the N_{yrs} (i.e., 36 for ERA-5) training years, we delete 7.5% of 36 = 2.7, which we round to 3 years. This is done N_{yrs} times, each time removing a different subset while ensuring that the deleted years are uniformly sampled, thereby avoiding that a certain year is recurrent in many of the SCMs while others are not. This procedure of removing a percentage is done multiple (N_{perc}) times, once for each percentage in the list *perc_yrs_out* (i.e., for ERA-5, the list of percentages are 5, 7.5, 10, 12.5, and 15; thus $N_{\mathrm{perc}} = 5$). Criterion (ii) is tested by redoing the previous step, but then the composite dates are shifted by n_{days} in time. These date shifts with respect to the composite dates are listed in param = *days_before*. For ERA-5, date shifts are 0, 7, and 14; thus $N_{\mathrm{shifts}} = 3$. In total, the subsampling leads to $N_{\mathrm{yrs}} \times N_{\mathrm{perc}} \times N_{\mathrm{shifts}} = N_{\mathrm{tot}}$ SCMs. For ERA-5 data, N_{tot} is equal to 540.

Next we calculate (and normalize) the frequency for each grid cell to obey criterion (i) and (ii), and we reject those that are not extracted at least 80% (param = *FSP_thres*) of the N_{tot} SCM maps. We found that using other reasonable parameter settings lead to qualitatively the same results. To form individual precursor regions, we use density-based spatial clustering of applications with noise (DBSCAN; Schubert et al. 2017), which assigns separate labels to groups of adjacent robust grid cells of the same sign (see Fig. C1 in appendix C). To achieve this, we use the Haversine formula as the distance metric, which calculates the great-circle distance between two points on a sphere.

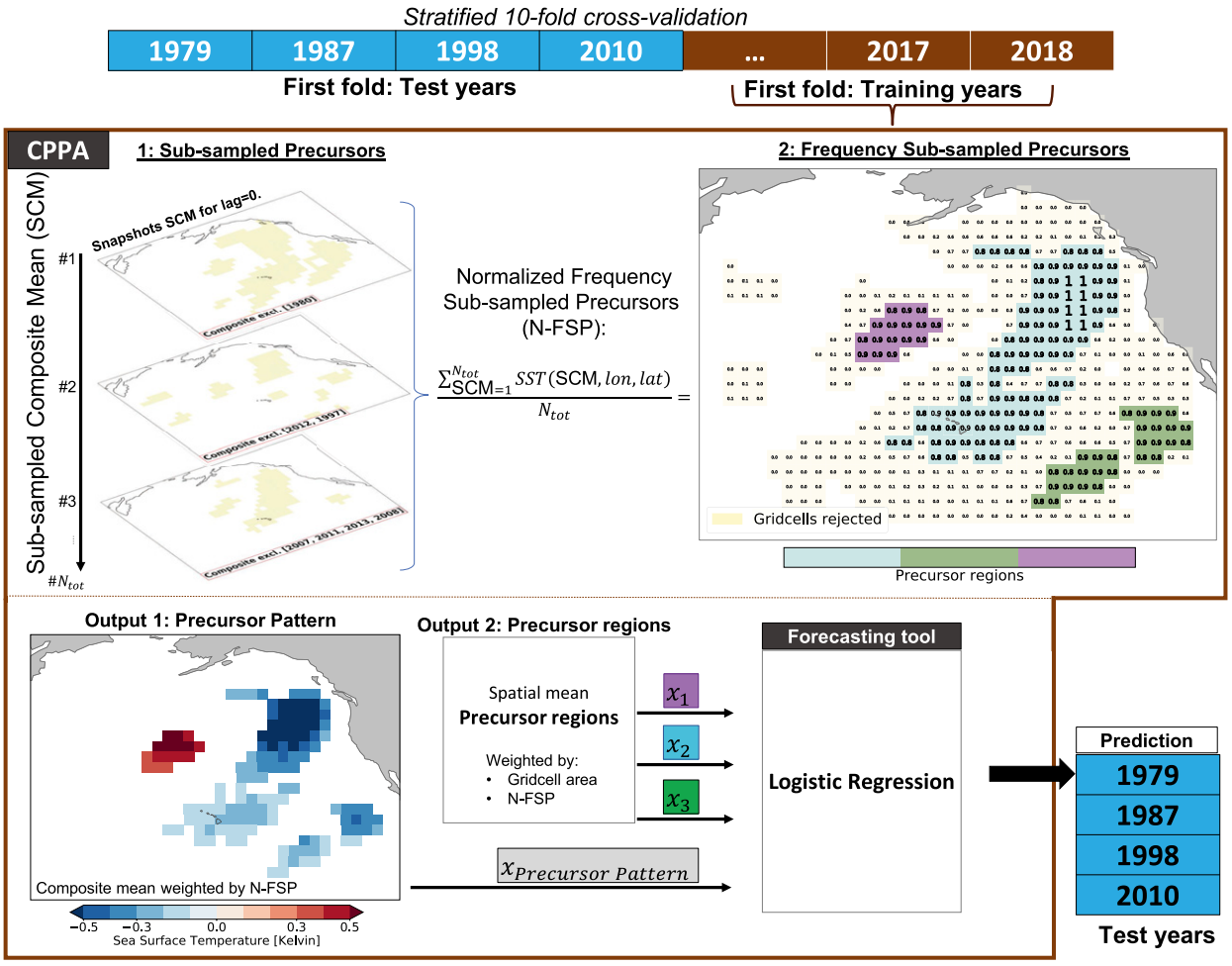

FIG. 1. Schematic illustration of the composite-based precursor algorithm (CPPA). In the upper left we illustrate step 1, which detects grid cells with robust composite anomalies for given lead time. In short, we define robustness by selecting those grid cells that consistently exceed a percentile threshold, irrespective of the subsample used (i.e., by leaving out some years or by shifting the composite lead times by 2 days). In step 2, we reject all nonrobust grid cells and the remaining (robust) grid cells are grouped into precursor regions (shown in different colors). See step 1 and step 2 in section 2c for more information. The output of the algorithm forms the predictors of the forecasting tool and consists of the spatial covariance of the full precursor pattern, and the spatial mean of all individual precursor regions (x_1, x_2, ...).

To summarize, CPPA searches for a robust SSTA pattern associated with the events of interest and using DBSCAN assigns separate labels to adjacent robust grid cells, thereby grouping them into precursor regions. Each of these precursor regions are reduced to a one-dimensional time series by calculating the area-weighted mean. Similar to Eq. (2), we also calculate a spatial covariance time series of all the precursor regions together, referred to as CPPA spatial pattern time series, or short CPPAsp. Hence, CPPA outputs both the spatial pattern time series *and* a single time series for each precursor region. For more detailed information about the output of the algorithm and a comparison to using the linear Pearson correlation metric, see appendix C.

d. Forecasting method

We implement a logistic regression (Varoquaux et al. 2015), which tunes the regularization coefficient using cross validation. Conventional logistic regression optimizes the coefficients to minimize the loss function, which tends to lead to overfitting. The regularization improves generalizability to unseen validation data by minimizing

TABLE 1. Parameters of the composite-based precursor pattern algorithm.

	Parameter names	Settings for ERA-5 data	Settings for EC-Earth data
1	*SCM_percentile_thres*	0.95	0.95
2	*perc_yrs_out*	5, 7.5, 10, 12.5, and 15	10, 20, 30, and 40
3	*FSP_thres*	0.80	0.80
4	*days_before*	0, 7, and 14 days	0, 7, and 14 days
4	*min_area_in_degrees2*	5^{o2}	5^{o2}
5	*distance_eps*	500 km	500 km

TABLE 2. Contingency table.

		Event observed	
		Yes	No
Forecast	Yes	True positive (tp)/hit (H)	False positive (fp)/false alarm (FA)
	No	False negative (fn)/misses (M)	True negative (tn)/correct negatives (CN)

the loss function + $1/C$ times the sum of the squared coefficients (L_2 *regularization*), with $1/C$ being the regularization coefficient. The $1/C$ is tuned by a second stratified fivefold cross validation (i.e., the model is trained on a subset of the data and subsequently, the generalizability of the model is tested on validation data). The regularization coefficient that renders the best average score on the 5 validation sets is chosen. Using the best value for C, the model is refitted on the complete training dataset (36 yr). We choose this statistical model because it does not have as many degrees of freedom as complex machine learning models and therefore is less prone to a limitation by data points.

Note that we are first separating the train-test split via stratified cross validation and subsequently split each training set into train-validation sets via another stratified cross validation. This allows us to efficiently use as much data as possible for training, while the test data are always strictly separated. For more information on this double cross-validation framework and a schematic overview, see appendix B. All precursor time series are standardized, where the mean and standard deviation are based on training data.

e. Forecast validation

According to Wilks (2011), ''forecast skill refers to the relative accuracy of a set of forecasts, with respect to some set of standard reference forecasts.'' A good quality forecast should meet a number of requirements, which cannot be summarized by a single scalar quantity (Wilks 2011). The World Meteorological Organization set up standard guidelines (WMO 2006) for verification of long-range forecasts, encouraging the use of relative operating characteristic (ROC), reliability curves, and a mean squared skill score (i.e., Brier skill score). We argue that one can only claim predictive skill if it performs well on all metrics. In addition, the forecast should perform better than an appropriate reference forecast, which for subseasonal predictions is the climatological probability. We use area under the curve relative operating characteristic (AUC-ROC) and area under curve precision-recall (AUC-PR), Brier skill score, and reliability plots.

The AUC-ROC was also used by McKinnon et al. (2016). The ROC represents a balance between true positive rate (TPR) and false positive rate (FPR) for different thresholds of the binary forecasting time series (Tables 2 and 3). The ROC area can be interpreted as ''the probability that the forecast probability assigned to the event is higher than to the nonevent'' (Mason and Graham 2002). See also Kharin and Zwiers (2003), Fawcett (2006), and Wilks (2011) for more information.

The AUC-ROC does not take into account the precision, reliability, and resolution of the forecast. Although the precision-recall curve still does not take into account the reliability and resolution, it is more suitable for imbalanced classes (Saito and Rehmsmeier 2015) and has a focus on evaluating the positive predictions (i.e., the forecast events). It quantifies the balance between precision and the Recall (or TPR) for different thresholds. If we forecast events using a low threshold, it is easy to get a very high precision, but difficult to get a high TPR (the denominator will be high due to many false negatives).

TABLE 3. Summary of verification metrics used in this article, see Table 2 for the contingency table.

	Calculation		Description
BSS	$(BS_f - BS_c)/BS_c$		Mean square error for binary classification (forecast vs climatology)
Precision	tp/(tp + fp)	$H/(H + FA)$	Correct positive predictions vs all positive predictions
Accuracy	(tp +tn)/(tp + tn + fn + fp)	$(H + CN)/(H + CN + M + FA)$	Ratio of total correct predictions
TPR (recall)	tp/(tp + fn)	$H/(H + M)$	Correct positive predictions vs total number of events
FPR or 1 − specificity)	fp/(fp + tn) or 1 − tn/(tn + fp)	$FA/(FA + M)$	Incorrect positive predictions vs incorrect positive predictions + correct negative predictions
AUC-ROC	Area under curve TPR vs FPR points		Forecast probability assigned to event higher than to nonevent
AUC-PR	Area under curve precision vs TPR points		Does not consider true negatives (misses); focuses on positive predictions

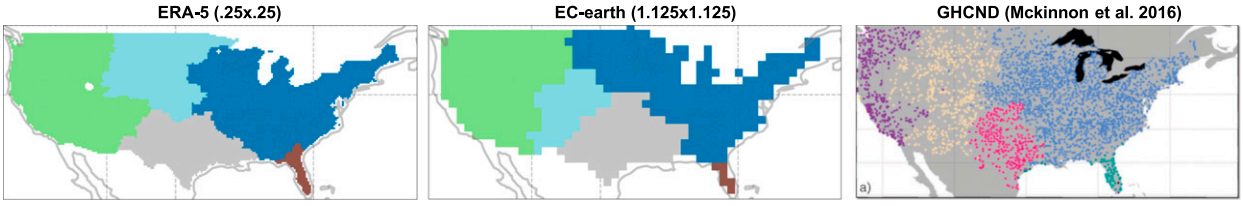

FIG. 2. Result of clustering each location on the basis of a binary time series, containing information on the timing of exceedances of a large anomaly (1 or 0). The datasets differ in spatial resolution (ERA-5: 0.25°; EC-Earth: 1.125°) and time period [ERA-5: 1979–2018; EC-Earth: 160 × 1 yr (present-day climate)]. The original clusters as presented in McKinnon et al. (2016) by using GHCND station data: 1980–2015.

The Brier skill score (BSS) is a commonly used metric for quantifying the quality of a probabilistic forecast. It takes into account both the reliability and resolution (Wilks 2011). The reliability quantifies to what extent forecast y_i deviate from the conditional average observation [mean of distribution of observations ($\overline{o_i}$), conditioned on the forecast (y_i), i.e., $\overline{o_i} = p(o_i|y_i)$]. Resolution quantifies the difference between the conditional average observation ($\overline{o_i}$) and the climatological probability ($\overline{o_i} - \overline{o}$) (i.e., forecasts with high resolution can more confidently distinguish events from nonevents). The BSS is calculated using the Brier score (BS) for a given probability time series:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2, \quad (0 \le \mathrm{BS} \le 1), \tag{3}$$

with p_i being the forecast probability at time step i and o_i being the observed event (1 or 0). The climatological Brier score BS_c is calculated for each train-test split by assuming a constant climatogical probability on the basis of the concomitant training datase (i.e., the same as used to fit the statistical model). Using BS_c and the Brier score of the forecast BS_f, we calculate the Brier skill score (if the BSS is significantly above 0, the forecast system is better than climatology):

$$\mathrm{BSS} = \frac{\mathrm{BS}_c - \mathrm{BS}_f}{\mathrm{BS}_c} = 1 - \frac{\mathrm{BS}_f}{\mathrm{BS}_c}, \quad (\mathrm{BSS} \le 1). \tag{4}$$

The reliability diagram (e.g., the last row of Fig. 6, described in more detail below) is used to visualize how reliable and resolute the forecast is. On the x axis we plot the forecast

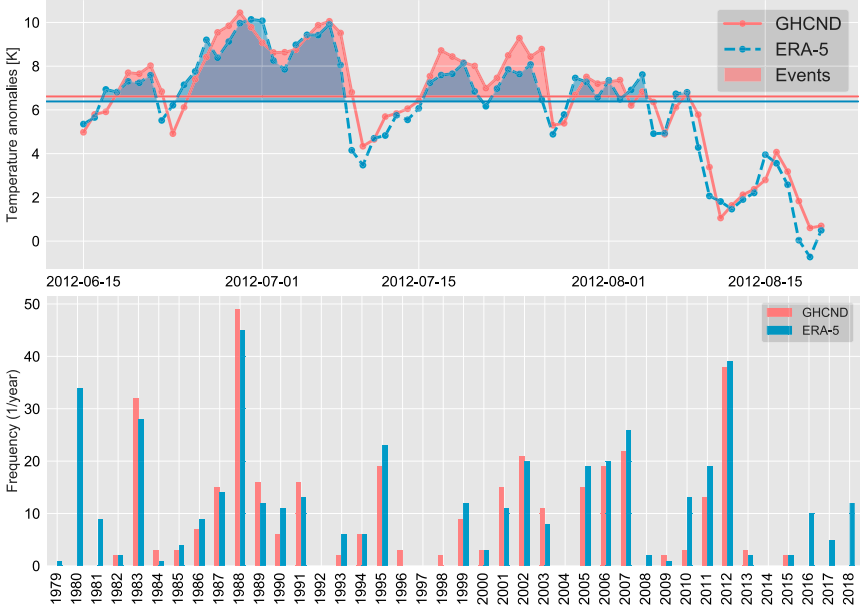

FIG. 3. (top) Year 2012 of original $T95$ time series and hot-day events based on observational GHCND station from McKinnon et al. (2016) and the $\mathrm{T90}_m$ time series based on ERA-5 reanalysis data. Events are defined as exceeding 1 standard deviation σ of $\mathrm{T90}_m$. (bottom) Frequency of hot days per year. For comparison in the top plot the mean and σ of both time series is calculated over the same period (1982–2015) as is available from the original $T95$ time series, for the bottom plot and elsewhere we use the whole ERA-5 time series (1979–2018).

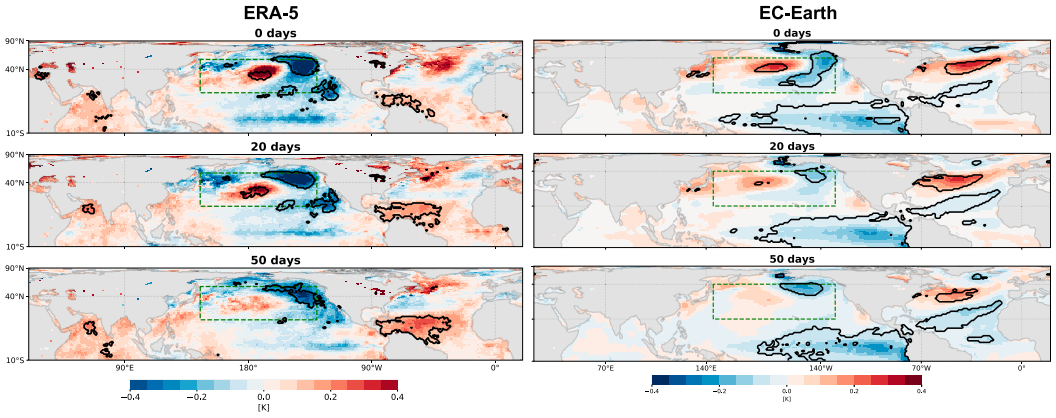

FIG. 4. Composite mean of hot-day events (mean over 10 training datasets presented) for both (left) ERA-5 and (right) EC-Earth. The lag with respect to hot-day events is presented in the subtitles. The stippled green rectangle depicts the PEP pattern. The contour lines show the robust anomalous grid cells that are extracted in at least 5 of 10 training datasets.

probability ranging from 0 to 1 (with 1 being 100% probability). For the reliability curve we use 10 equally sized bins (step size = 0.1) and plot the forecast probability on the x axis and observed frequency on the y axis. A perfectly reliable probabilistic forecast would always match the observed frequency (i.e., show a diagonal line). A histogram is plotted below the curve to show the forecast distribution, which informs about the sharpness of the model. The sharpness refers to the ability of the forecast model to substantially deviate from the climatological probability. The dark-gray area shows where the forecast is better than climatology (BSS $>$ 0), and the light-gray area shows where the forecast is only doing better than a random forecast.

Confidence intervals are created by bootstrapping (n = 2000, unless stated otherwise), where we bootstrap blocks to account for autocorrelation, thereby avoiding oversampling of dependent data points. The block size is objectively defined by the lag at which the autocorrelation of the target becomes significantly different from zero.

3. Results

a. *Spatial clustering and hot days in ERA-5 and EC-Earth*

We performed a parameter sweep to test for robustness of the eastern U.S. cluster in the ERA-5 and EC-Earth datasets, as further detailed in appendix A. Overall, we conclude that the eastern U.S. cluster is robust [i.e., it is generally categorized as a separate cluster, with only small differences in the exact boundaries and size (depending on minor perturbation of the clustering parameters)].

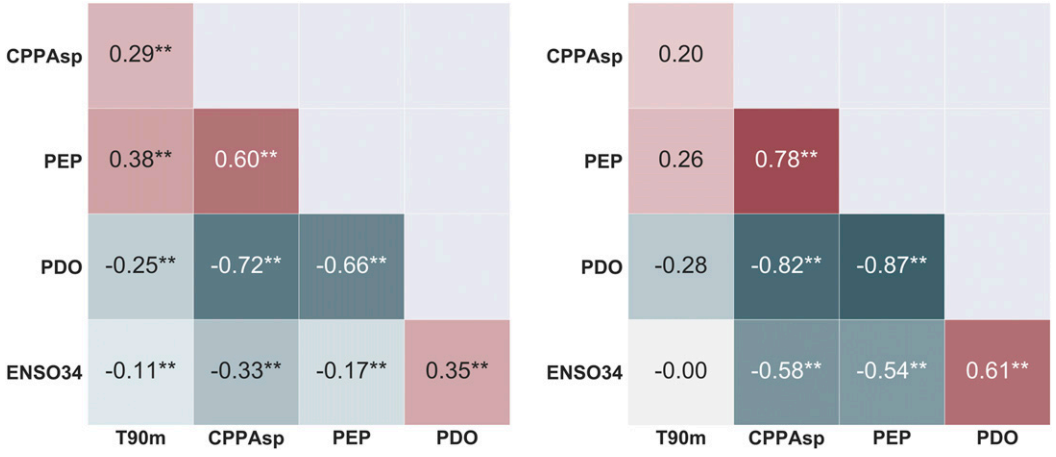

FIG. 5. (a) Out-of-sample cross-correlation matrix for daily data during the study period (24 Jun–22 Aug) and for (b) annual mean values. The ENSO time series refers to the Niño-3.4 time series, defined by the area-weighted SSTA mean between 5°S and 5°N and between 170° and 120°W. The plots are based on ERA-5 data. The double asterisk indicates significance at p value $<$ 0.01.

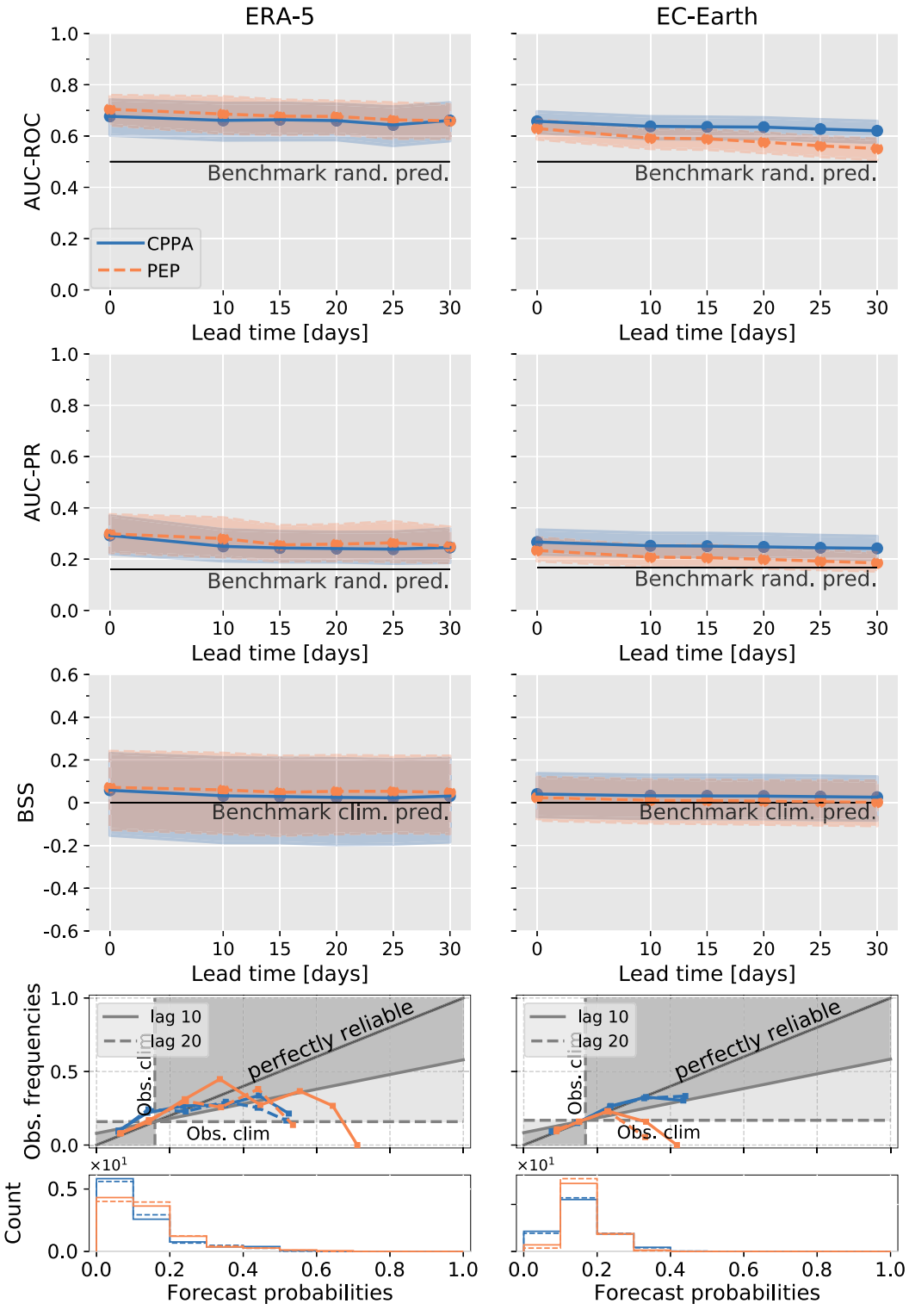

FIG. 6. Forecast validation for hot days, using only information from SSTA. We compare using the PEP pattern with the CPPA precursors for forecasting and show the importance of using multiple skill metrics.

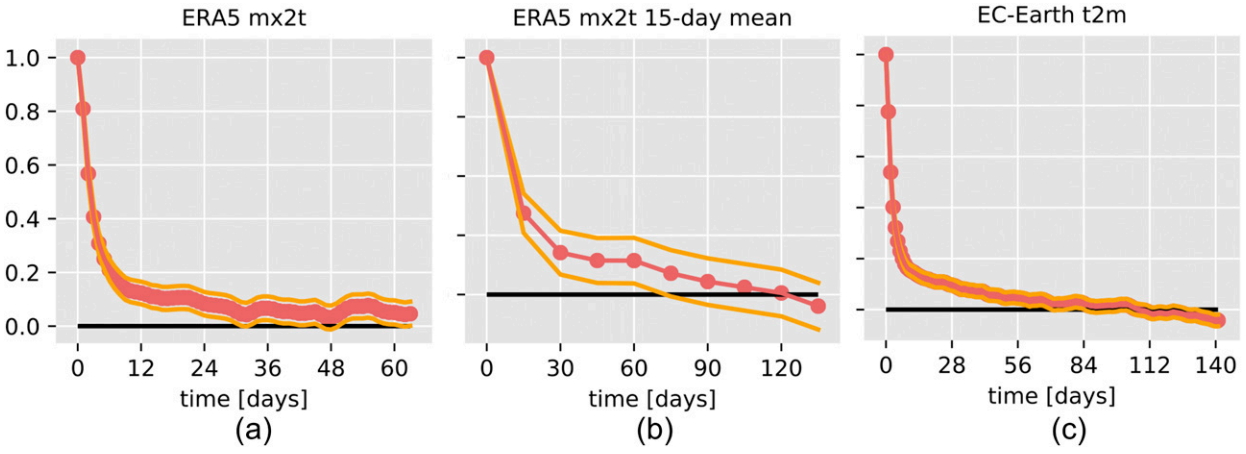

FIG. 7. The autocorrelation of $T90_m$ and the 15-day mean $T90_m$ in ERA-5 and the $T90_m$ in EC-Earth. The autocorrelation is used to determine block window size for bootstrapping.

As described in appendix A, we choose the eastern U.S. cluster that is most similar to McKinnon et al. (2016), see Fig. 2. We calculate the spatial $T90_m$ and the associated hot days as explained in section 2b. Figure 3 shows that the ERA-5 $T90_m$ time series and associated frequency of hot days (yr^{-1}) matches closely with the original T95 and hot-day time series found by McKinnon et al. (2016).

b. Comparison between CPPA, PEP, and climate indices

Figure 4 shows the hot-day events composite mean of SST grid cells (mean over 10 training datasets) for both ERA-5 and EC-Earth, where the stippled green rectangle depicts the PEP pattern and the black contour lines show the robust anomalous grid cells detected by CPPA. As can be seen from Fig. 3, there is a lot of interannual variability in the amount of hot days, with 4 years together accounting for 33% of the events and with 9 years having less than 1% of the events in the ERA-5 reanalysis. The output of CPPA, however, is robust across the 10 training datasets, as detailed in appendix B (Fig. B1). For ERA-5, the labels that are (randomly) assigned to each precursor region by the DBSCAN clustering algorithm are shown in appendix C (Fig. C1).

We observe that, in the tropical Pacific, a La Niña–like pattern is picked up in EC-Earth and not in ERA-5 and also the tropical Atlantic precursor regions are different. Both ERA-5 and EC-Earth do share the cold-eastern and warm mid-Pacific features. These are also the main features of the PDO pattern and are part of the PEP pattern as presented by McKinnon et al. (2016). Yet the cold western Pacific of the PEP pattern is considered nonrobust according to CPPA.

We also analyzed how the $T90_m$, PEP, Niño-3.4, PDO, and CPPA spatial pattern (CPPAsp) time series are linked to each other via a cross-correlation matrix (Fig. 5). See appendix E for background information on the calculation of the PDO and ENSO indices. We observe that the PEP time series show a higher correlation coefficient with $T90_m$ when compared with the CPPAsp time series and the climate indices (PDO and Niño-3.4), particularly during the summer days. We also observe that CPPAsp and PEP are strongly correlated with the PDO time series. The difference between EC-Earth and ERA-5; the link between PEP, CPPAsp, and the climate indices; and the potential physical mechanism are further discussed in section 4c. In the following section, we will compare the forecast skill between PEP, the climate indices, and the CPPA time series (CPPAsp and CPPA precursor regions time series).

c. Using multiple validation metrics

Figure 6 shows the verification of hot-day-event forecasts, comparing the use of the PEP time series versus the CPPA output to fit the statistical model. As explained in section 2e, we objectively determine the block window size for bootstrapping by calculating up to which lag the autocorrelation is significantly different from 0 (see Fig. 7), for the ERA-5 daily $T90_m$ time series this is 32 days (Fig. 7a) and for the EC-Earth daily $T90_m$ time series this is 71 days (Fig. 7c).

We observe that forecasts based on either PEP or CPPA perform better than random chance, rendering approximately the same skill for ERA-5. For EC-Earth data, we observe that CPPA is a better precursor compared to PEP. We also see lower skill for EC-Earth, even though the climate model data has 4 times as many data points. Both datasets, however, do not render a significantly better forecast compared to the climatological probability, as is evident from the near-zero BSS values and the reliability diagrams (Fig. 6). This nonexistent predictability for hot days is not surprising given the fact that we are trying to predict the exact day at which the hot-day event should occur. Even for the EC-Earth data, where we have many data points available, the statistical model cannot resolutely discriminate between events and nonevents. Since we know the EC-Earth model has its limitations in representing the real climate, especially extremes, we will now only focus on the ERA-5 dataset.

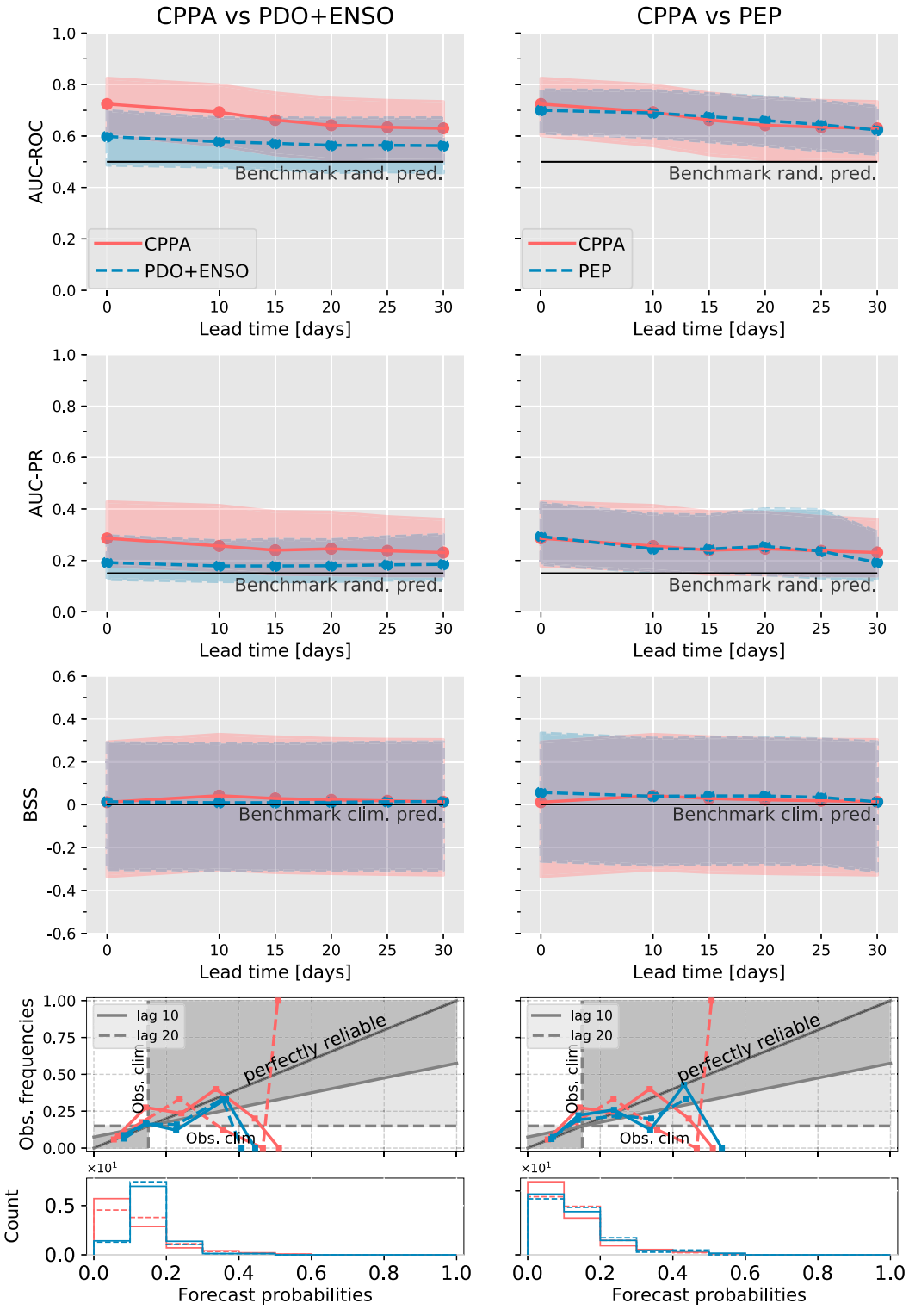

FIG. 8. Forecast validation for "hot 15-day mean events." Here we show the comparison between using (left) the PDO + ENSO vs the CPPA precursors and (right) the PEP pattern vs the CPPA precursors.

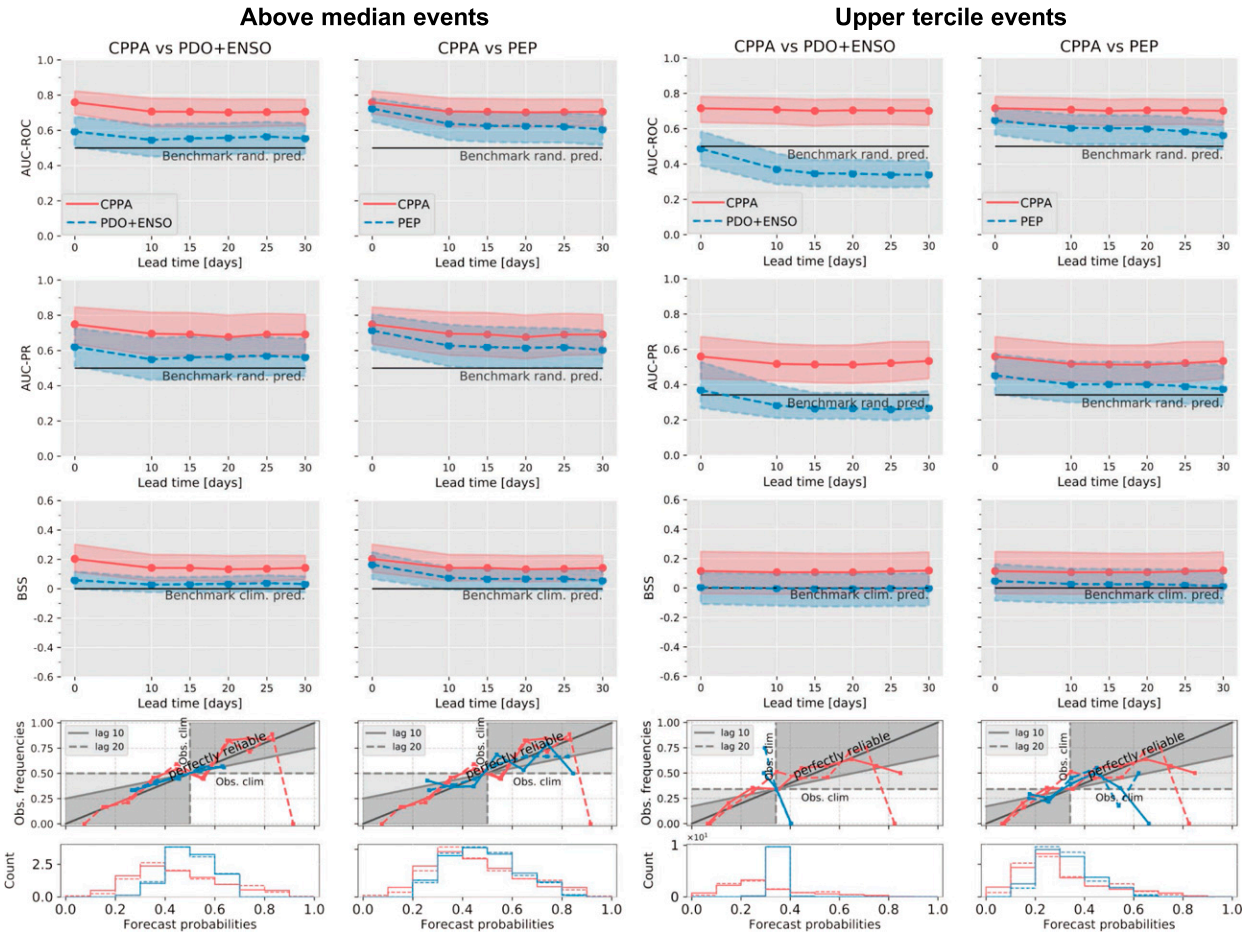

FIG. 9. As in Fig. 8, but (left),(left center) for above-median events and (right center),(right) for upper-tercile events of the 15-day mean $T90_m$ time series. The plots are based on ERA-5 data.

d. Temporal aggregation to improve signal-to-noise ratio

To improve the signal-to-noise ratio, we aggregate over time with the trade-off of a reduction in temporal precision and the number of data points. We aggregate the daily data into bins of 15 days and calculate the mean of all bins. The window size of 15 days is commonly used in the literature when studying Rossby wave dynamics (Kornhuber et al. 2017; Röthlisberger et al. 2018). Since we are now working with time windows, the lead time is defined from the day that the forecast would be issued, using only information prior and including that day, to the center date of a forecast time window, see appendix F for more information. We will compare the forecasts with the conventional approach [i.e., using the relevant climate modes of variability from SST (PDO + ENSO)].

Figure 8 shows the verification when we first calculated 15-day means of $T90_m$ and then used the event definition that was also used to define hot days [see Eq. (1)], thus having a base rate of approximately 16%. The block window size is five time steps (i.e., 75 days). For these so-called hot 15-day mean events, we observe a decline in skill, not an improvement. The histogram shows that almost all values are close to the climatological probability, especially for the PDO + ENSO forecast. Ostensibly, we still have insufficient information to fit a reliable model and/or the reduction in data points seems to dominate the benefit of a better signal-to-noise ratio.

Thus, next, we lower the extremity of the events (which increases the base rate) and define the target based on 15-day upper-tercile and 15-day above-median events. Figure 9 shows that the statistical models that are fitted using the CPPA precursors outperform the ones that use PEP, or PDO + ENSO. For the upper-tercile events (right two columns of Fig. 9), skill is better relative to the hot 15-day mean events (Fig. 8) but is slightly lower relative to the above-median events, which show skill up to at least 30 days of lead time (left two columns of Fig. 9). To summarize, we improved forecast skill by 1) finding better precursors using CPPA and 2) using temporal aggregation in combination with increasing the base rate (i.e., lowering the threshold for events).

e. Using a window probability and spatial aggregation to improve event forecasts

To increase predictability of extreme events, we relax the temporal precision by using a ''window probability,''

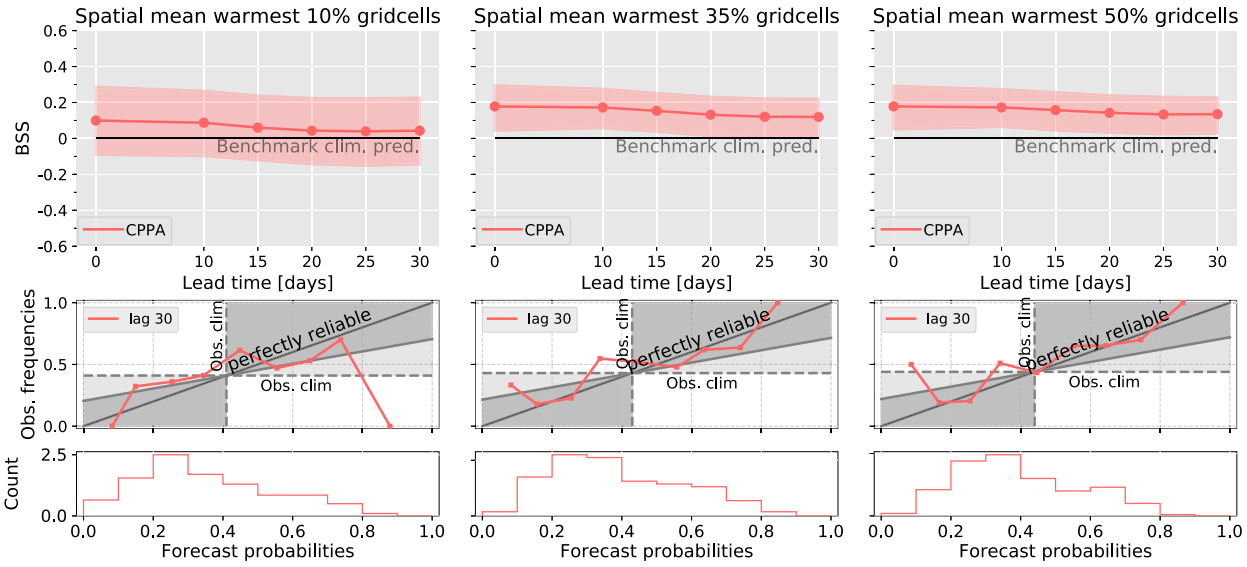

FIG. 10. Forecasting heat waves (defined in the text) within a 15-day window. Three different spatial aggregation sizes are used to define our continuous temperature time series [(left) $T90_m$, (center) $T65_m$, and (right) $T50_m$), after which the associated moderate heat-wave events are calculated. The plots are based on ERA-5 data.

meaning that we predict the occurrence of a relatively short heat-wave event within a longer time window. Hence, the exact date of occurrence within this time window is flexible. When using a 15-day time window, a predicted heat-wave event may thus occur 7 days earlier or later. We define a heat wave when two or more hot days occur with at most one nonhot day between them. With this approach, we still smooth out noise in the *precursor* time series (by using 15-day means), while still predicting relatively short-lived events consisting of daily temperature extremes.

Still, the target variable is not smoothed in time. To increase the signal-to-noise of the target variable we apply spatial aggregation. We do this in a similar manner as was done for $T90_m$ [section 2b(1)], defined as the *spatial* mean over the 10% warmest grid cells within the eastern U.S. cluster. Here, we define two additional target time series with increased spatial aggregation by calculating the mean over the 35% warmest ($T65_m$) and 50% ($T50_m$) warmest grid cells. Subsequently, the hot days are defined for each time series using the equivalent of Eq. (1).

The Brier skill score for the $T90_m$ heat-wave forecast (Fig. 10, left column) is lower relative to that of upper-tercile 15-day mean $T90_m$ events (Fig. 9, right columns), even though the base-rate of the $T90_m$ heat-wave window probability is higher (41%). By aggregating over space ($T65_m$ and $T50_m$, i.e., the second and third column of Fig. 10) one reduces the noise in the target time series and thereby enhances forecast skill.

f. *Subseasonal forecasts of moderate heat waves using both SST and soil moisture*

Previous results focused on quantifying predictability from only SST. Now we aim at enhancing forecast skill by including additional information from soil moisture. We proceed with $T65_m$ as it has significant skill up to at least 30 days (Fig. 10, central column), while still being relevant for temperature extremes. During the summer days, the daily $T65_m$ time series[1] has a temporal mean value of 2.3°C and standard deviation of 2.1°C. Using the equivalent of Eq. (1), the events have an average anomaly of 5.6°C ranging between 4.4° and 9.9°C. A total of 466 days belong to these events (base rate of 15.5%), and after grouping these days into multiday events (as defined in section 3e) there are 103 events left. Because the threshold is now less extreme, we will call these events moderate heat waves.

Figure 11 shows the verification results for forecasts when using precursors both from CPPA and soil moisture (orange dashed line) and when using only CPPA (blue solid line). We observe that soil moisture contributes to a small increase in skill up to 30 days lead-time, but for longer lead times all information can be retrieved from the SST precursors. [Tables C1 and D1 in appendixes C and D, respectively, show all precursors that were used for this prediction. In appendix F, Fig. F2 shows that the 10 models with a lead time of 50 days that were learned on the basis of different training datasets are robust (i.e., the 10 models generally learned the same regression coefficients). Figure F3 shows that also the forecast quality is robust when using different train-validation combinations; see also appendix B.]

[1] Note that $T65_m$ refers to a time series from calculating the spatial mean of the 35% warmest eastern U.S. grid cells on each day; it has a temporal mean and standard deviation (just like $T90_m$ is a time series, as shown in Fig. 3).

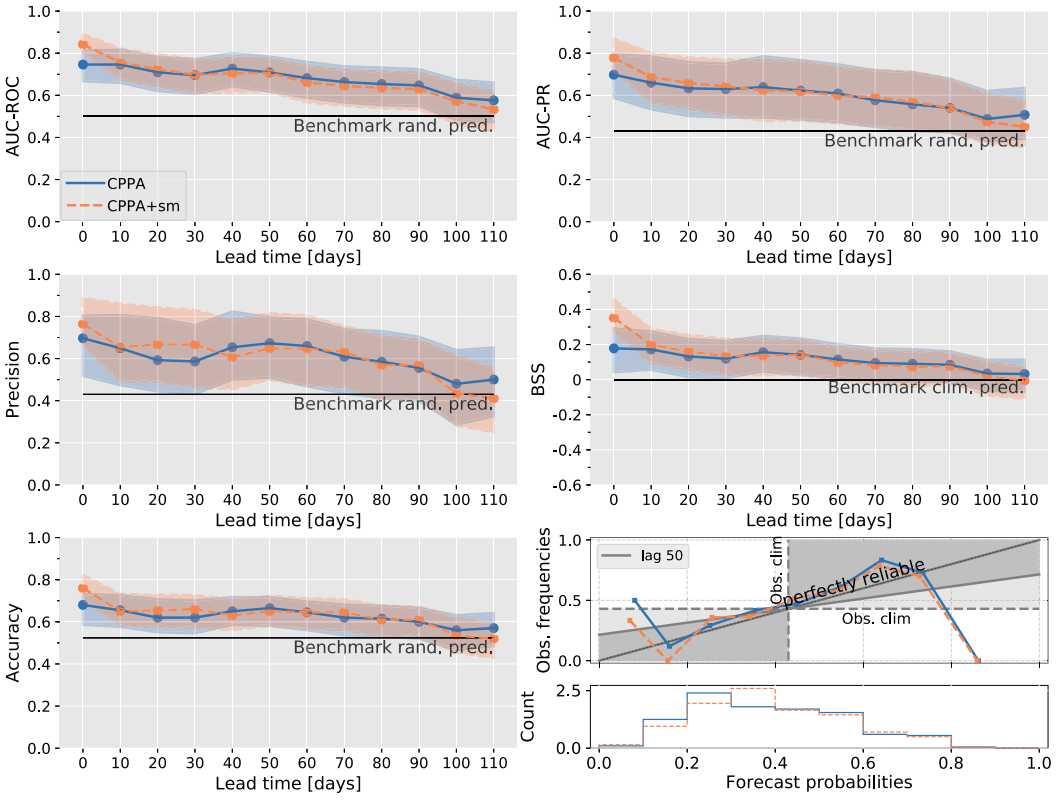

FIG. 11. Verification results for forecasting T65$_m$ heat waves (defined in the text) within a 15-day window. Solid blue line shows the results for the forecast when using the CPPA time series, and the stippled orange line show results when including precursor time series from both CPPA and soil moisture (see Tables C1 and D1 of appendixes C and D, respectively, for a list of all precursors that were used). Bootstrap sample size is 5000. The plots are based on ERA-5 data.

For this forecast, we achieve predictive skill 50 days in advance at the 2.5th–97.5th confidence interval (n = 5000). This forecast for moderate heat waves is more capable of discriminating between event and nonevent occurrences (higher resolution) compared to the original hot-day definition, as is evident from the reliability diagram and Brier Skill Score.

4. Discussion

a. *Using multiple validation metrics*

A proper forecast validation for eastern U.S. hot days (i.e., forecasting individual days), shows that the forecast does not perform significantly better than the climatological probability. The probabilistic forecast values for hot days were not able to confidently discriminate between events and nonevents [i.e., low resolution, $p(o_j|y_i)$, where y_i is the forecast probability, and o_j are the observed values (Wilks 2011)]. This can be seen from the reliability diagrams in Fig. 6. Contrarily to McKinnon et al. (2016), we conclude that there is no predictive skill for individual hot days.

The AUC-ROC metric measures discrimination (see section 2e), also the forecast values are only sorted and their actual value is neglected. Thus, resolution will not be measured, and consequently the forecast probability might be always close to the climatological probability (p_c), which makes them of low practical value. If one wants to assess predictive skill, the AUC-ROC is an improper validation metric if used by itself as it only measures *potential* skill (see Wilks 2011, chapter 8).

b. *Improving statistical forecasts for events*

The problem when predicting extreme events on S2S time scales lies between a boundary condition problem and an initial value problem (Vitart et al. 2019), that is, the boundary conditions that we use to constrain a target distribution (in this case temperature) changes over time. From this perspective, we believe there are three limiting factors for these statistical forecasts: 1) missing information of low-frequency drivers, 2) the chaotic nature of the atmosphere [i.e., knowing the full constrain of the boundary condition(s) is still not strong enough to reliably predict extreme events (Krishnamurthy 2019; Vitart et al. 2019)], and 3) the statistical model is suboptimal due to insufficient data points and/or the complex nonlinear interactions cannot be described by the model.

When using only PDO and ENSO for forecasting (see Fig. 9), we clearly miss information compared to using the

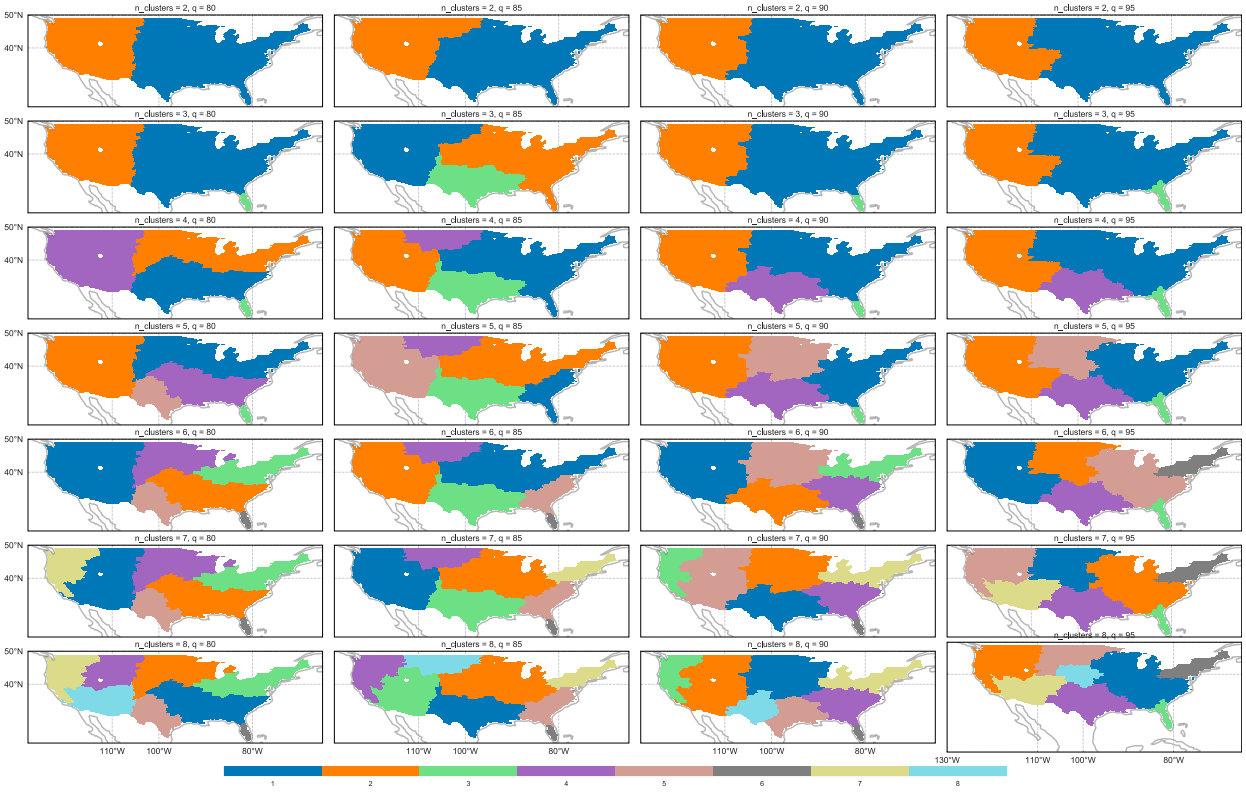

FIG. A1. Parameter-sweep spatial clustering results for ERA5 (0.25° × 25°).

CPPA time series. While we have many data points when using the EC-Earth dataset for the forecast of individual hot-day events (Fig. 6) or ''hot 15-day mean events'' (see Fig. F4 in appendix F), we are still unable to produce reliable and resolute forecasts. However, improved precursors (using CPPA) and temporal aggregation (relating to point 1 of the limiting factors) can only contribute to a pronounced improvement in skill when we predict events that are not too extreme (relating to point 2) (cf. Figs. 8, 9). We do not think the linear statistical model (point 3) was inadequate, we have tried tuning a tree-based gradient boosting regressor (GBR) by doing an extensive parameter grid search (results not shown). However, the best performance of the GBR was only as good as the regularized logistic regression.

To enable forecasts of more extreme events, we use a window probability definition for the target variable (i.e., the probability of a relatively short-lived heat wave occurring within a longer prediction window) and, in addition, apply stronger spatial smoothing. This combination effectively reduces the noise in the target time series and increases the base rate, while still predicting societally relevant high-temperature events. Thus, we conclude that S2S predictions of high-temperature events are possible, but also fundamentally limited by the chaotic nature of the atmosphere constraining the signal-to-noise ratio and the availability of data, which hampers the detection of the signal. Nevertheless, with the techniques presented here, a stakeholder can be helped to decide on the preferred balance between spatial aggregation, temporal aggregation, and extremity of the to-be-forecast events. Thus, given the stakeholder needs, optimal aggregation and threshold levels can be found to attempt to render skill at the desired lead times.

c. *Physical interpretation of the CPPA pattern*

In a response-guided approach the features are learned objectively, which can improve forecast skill relative to using, for example, climate indices, as is shown in this paper. Another important advantage is that the features remain physically interpretable, hence, they can be evaluated with physical understanding. Both ERA-5 and EC-Earth render a SSTA pattern that strongly resembles the main features of the PDO pattern in its negative phase (see Fig. E1 in appendix E). This is in line with the physical mechanism that low-level heating can effect the position of the jet stream (Thomson and Vallis 2018; Teng et al. 2019).

In the Atlantic Ocean, the relationship between hot days and SSTA differs between EC-earth and ERA-5. We suspect EC-Earth to suffer from biases, since model perturbation experiments have shown a reduction in precipitation due to a warm Gulf of Mexico state (Wang et al. 2010), which overlaps with the warm Caribbean Sea region

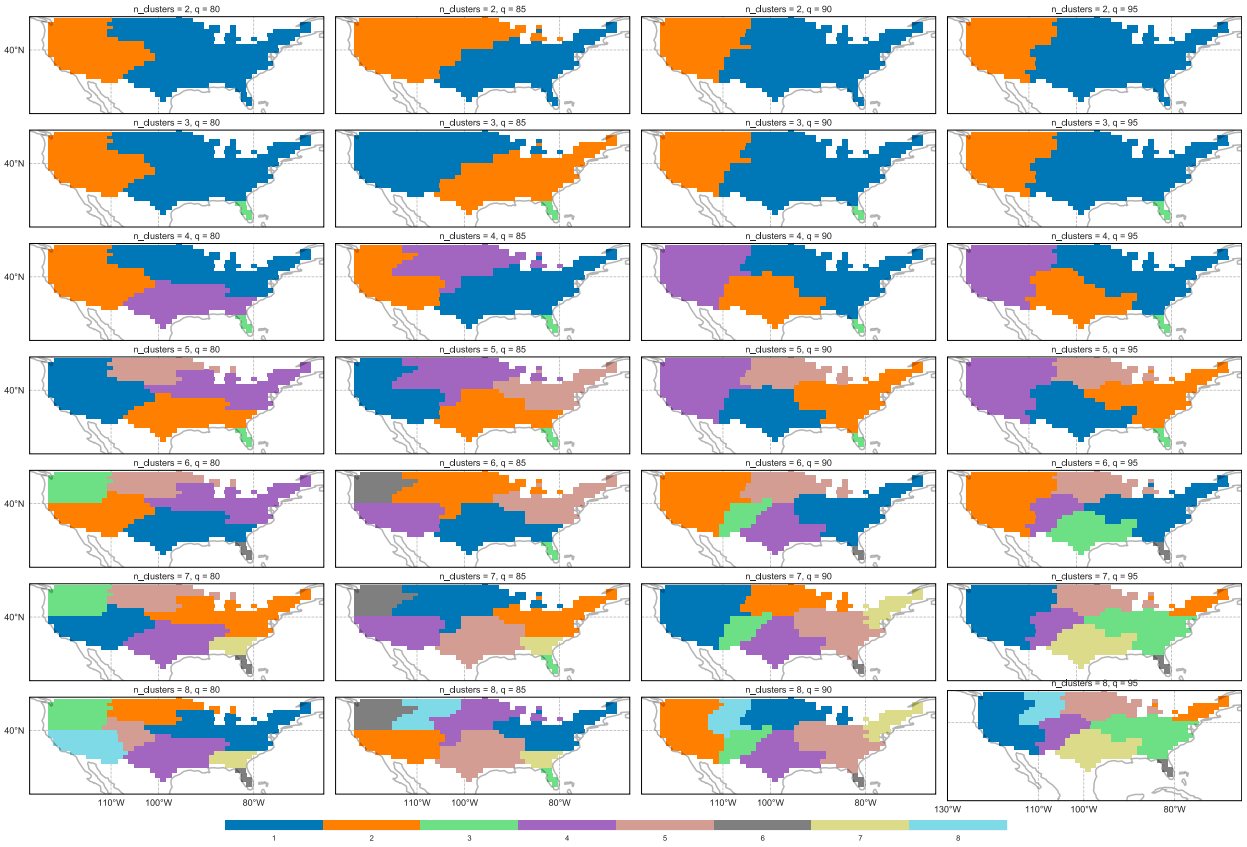

FIG. A2. Parameter-sweep spatial clustering results for EC-Earth (1.125° × 1.125°).

our analysis finds in the ERA-5 data (Fig. 4, left column). The lower amount of precipitation is linked to an increase in temperature due to a stronger soil moisture–temperature feedback (Wang et al. 2010). Their analysis describes the complexity of the physical links, indicating that it is difficult to simulate the teleconnection between U.S. temperature and Atlantic SSTA. EC-Earth has to simulate the entire chain of interactions accurately to get the correct temperature impact, e.g., the circulation, cloud and precipitation response, land surface fluxes and the soil–moisture temperature feedback.

In general, we also observe that the pattern anomalies are stronger for the ERA-5 dataset, which could be due to

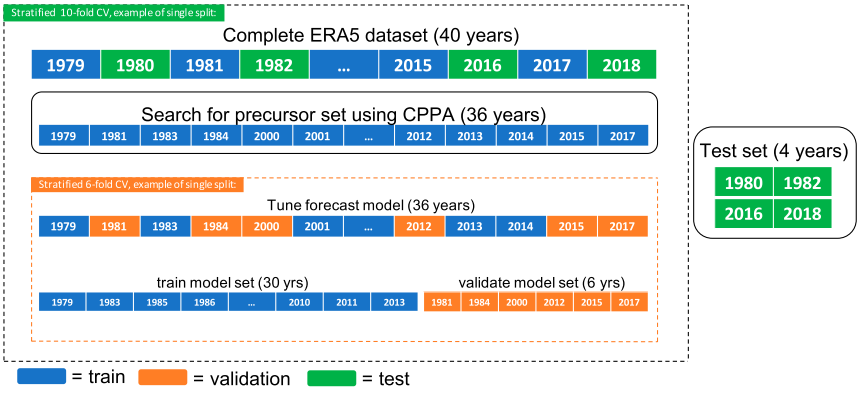

FIG. B1. A complete overview of the ''double'' stratified cross-validation procedure to enable a response-guided search for precursors and model tuning with a limited amount of data. This results in a forecast model for each 10 train-test splits and for each lag.

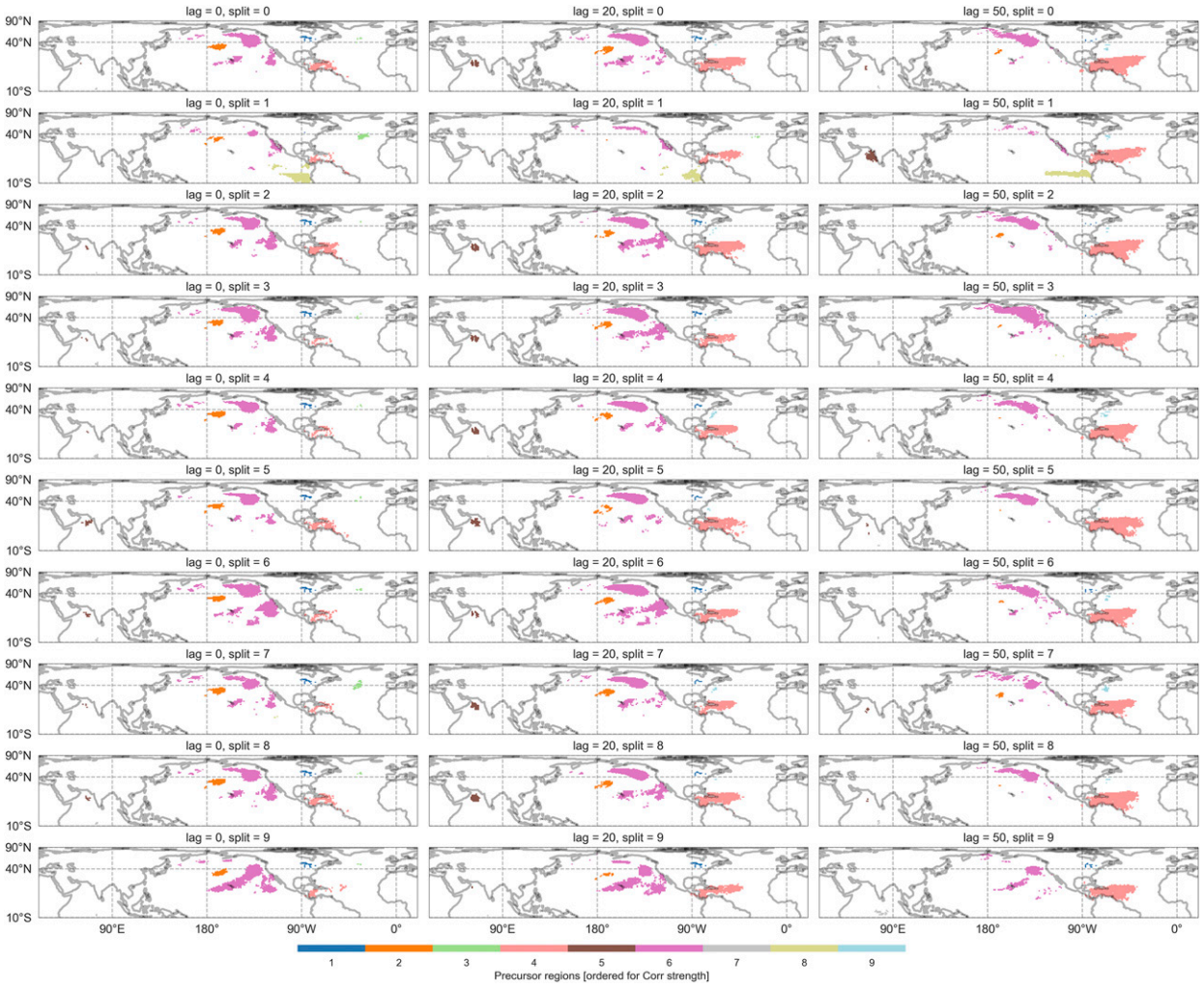

FIG. C1. Sea surface temperature regions found by the CPPA algorithm using a single training set (36 years). Clusters should be at least 5° by 5° big (defined at 45°N) to form a core sample; if they show a high density, they are more likely to include neighboring grid cells into the cluster. The radius at which core samples (initial clusters) search for neighboring grid cells is set by the *eps* parameter, in our case 500 km. We take into account the gridcell area by assigning weights to the samples (i.e., grid cells). Time series are calculated by taking daily means, weighted by gridcell area and the N-FSP; see section 2c and Fig. 1.

the sampling size of 40 years. However, we suspect it is more likely that EC-Earth is underestimating the link between SSTA and hot days, which is also supported by the lower forecast skill of EC-Earth presented in section 3c. The ostensible underestimation of the atmospheric response to SSTA could be the result of unresolved smaller-scale processes due to insufficient spatial resolution in climate models (Hodson et al. 2010; van Der Linden et al. 2019; Thomson and Vallis 2018).

McKinnon et al. (2016) proposed that the PEP pattern arises from atmosphere-to-ocean heat fluxes in spring/summer, which are indeed directed toward strengthening of the pattern (see Fig. S12 in McKinnon et al. 2016). This suggests a mechanism acting on a subseasonal time scale, separate from the PDO. However, using annual mean values, the cross-correlation matrix based upon ERA-5 data in Fig. 5 shows high correlation coefficients between the PEP and PDO, suggesting that PEP does not arise in a 60-day window, but is in fact, strongly related to the presence of the negative PDO phase.

We propose that the presence of the right background SSTA pattern favors the occurrence and persistence of a wavy jet stream resulting in a high pressure system over the eastern United States, and ocean–atmosphere heat fluxes are likely amplifying the final response (a wavy jet stream). The correlation of the SSTA pattern with temperature is likely strongest in summer (Fig. 5) because the impact of a high pressure system on temperature is exacerbated by the higher solar irradiation and potentially stronger soil moisture–temperature feedbacks (when the evaporation becomes strongly limited by the available soil moisture, the impact on temperature becomes most apparent, which

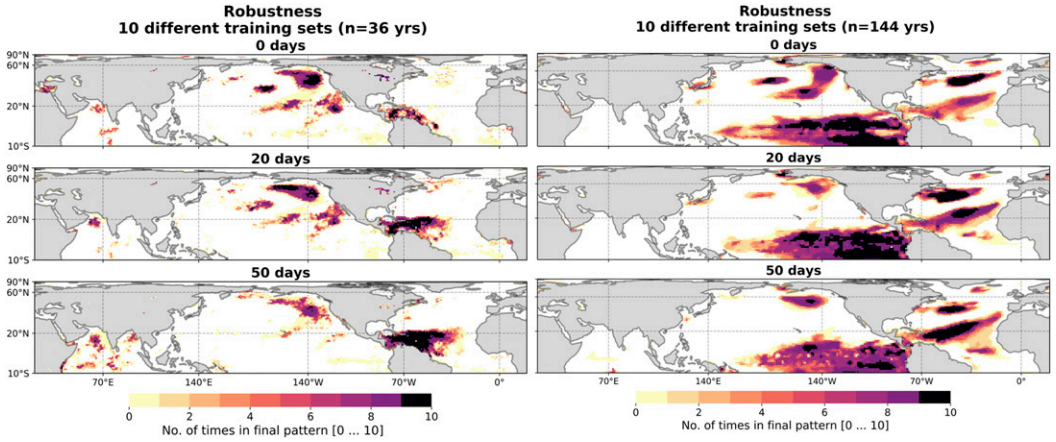

FIG. C2. Robustness of grid cells for (left) ERA-5 and (right) EC-Earth; see section 2 for details. Values equal to 10 mean that the grid cell is extracted in all 10 different training sets. Grid cells that are consistently part of the precursor pattern are interpreted as more robust.

generally happens at the end of the summer) (Seneviratne et al. 2010).

5. Conclusions

In this work, we focused on 1) a comparison between the response-guided CPPA approach, the PEP pattern, and using climate indices, 2) the importance of using multiple skill metrics and 3) how one can make reliable statistical S2S forecast for high-temperature events. First, we presented an algorithm that objectively extracts robust SST anomalies (SSTA) from a target event time series. We conclude that CPPA can successfully detect robust SSTA regions. We note that, using continuous time series instead of a binary one for the target variable, correlation maps appear more robust and render similar results (see discussion in appendix C). Boschat et al. (2016) also concluded that correlation maps are more robust than a composite approach, although they did not perform a subsampling as done by the CPPA to check for robustness.

The use of the AUC-ROC score as a single metric to assess skill should be avoided because it measures only potential skill. Based on multiple skill metrics, we showed that long-lead predictability does not exist for *individual* hot days (section 3c). To generate reliable S2S forecasts, one needs to improve the signal-to-noise ratio, either by temporal aggregation, spatial aggregation or statistical filtering techniques [e.g., wavelet transformations (Deo et al. 2017)]. Here, we have shown that a low signal-to-noise ratio in the target time series is indeed a bottleneck when trying to forecast extreme events defined on a daily resolution. Using a window probability, we were able to forecast moderate heat waves with an average anomaly of 5.6°C above climatology.

Forecast skill improved when using the CPPA precursor regions as compared with using modes of variability (PDO, Niño-3.4). A key advantage of this response-guided approach compared to some other feature extraction techniques, is that precursors remain physically interpretable. With this approach, one can benefit from a data-driven tool to optimize skill and also use physical understanding to e.g., identify plausible physical relationships, select variables, estimate the associated time scale of dynamics, understand limitations of predictability from physics (Mariotti et al. 2020). Hence, we recommend a response-guided approach to learning one's input features for statistical forecasting models, as was also done by Kretschmer et al. (2017), for which Python computer code is being developed and shared on Github.[2] The Github release contains the code and ERA-5 time series to reproduce the forecasts in this paper.

Our findings highlight how to get an improved physical understanding and more skillful statistical S2S forecasts by 1) objectively searching for precursors instead of using modes of variability and 2) improving the signal-to-noise ratio. Additionally, we introduced a window probability to allow temporal flexibility, which results in more reliable predictions of events compared to trying to predict the exact date of occurrence. A stakeholder is helped more with a *skillful* forecast with some uncertainty in exactly when the event will happen (e.g., between 48 and 62 days from now), as compared to an uncertain and *unskillful* forecast, which attempts to predict exactly when an event will happen.

Future work could look into implementing statistical methods to obtain a better signal-to-noise ratio. Using an automated response-guided approach as presented here in combination with dynamical model output (i.e., producing

[2] The code that was used for this work is published in a separate release (https://github.com/semvijverberg/RGCPD/releases/tag/v3.0.0). The most recent version is also online (https://github.com/semvijverberg/RGCPD).

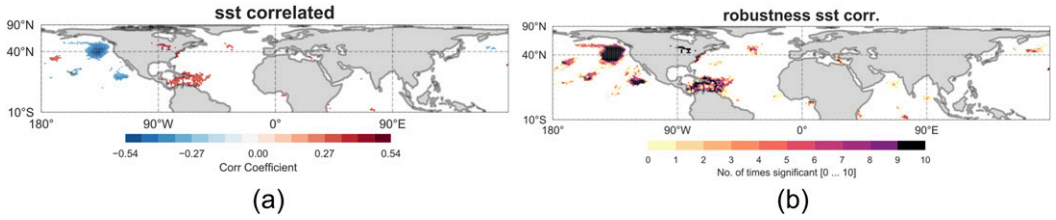

FIG. C3. (a) SST correlation maps ($\alpha = 0.01$) for 15-day mean time series at lag = 0, and (b) the robustness across different training sets. In (a), the mean is over training sets, and grid cells are masked if they were not in 50% of the training sets.

hybrid forecasts) could be the next step to make operational, improved S2S forecasts.

From a physical perspective, the link between SSTA and temperature is complex and appears to be affected by 1) the soil moisture–temperature feedback, 2) ocean–atmosphere interaction leading to a feedback between Rossby waves and the SSTA, 3) the direct circulation response to the SSTA pattern excluding the effect of ocean–atmosphere feedbacks, and potentially 4) a dependence of the atmospheric response on the wind field (Thomson and Vallis 2018). The physical interaction and relative importances of these processes will be subject of future work.

Acknowledgments. Sem Vijverberg thanks Anais Couasnon for help with phrasing the results, Bram Kraaijeveld for his earlier work on the forecasting part, and the two anonymous reviewers who gave good feedback and suggestions.

APPENDIX A

Spatial Clustering of Heat Extremes

A binary time series of extreme temperature event occurrences is calculated for each geographical location, the binary time series is 1 if the temperature exceeds the qth percentile and is 0 otherwise. The resulting strings of 0s and 1s are the input for the clustering algorithm. The binary strings that are very similar (i.e., those that experience heat extremes simultaneously) are clustered together. To be consistent with McKinnon et al. (2016), we use the hierarchical agglomerative clustering algorithm (Murtagh and Contreras 2012), with the ''jaccard'' distance metric (Jaccard 1912) and the linkage criterion is set to ''average,'' meaning that the average distance between the binary strings is minimized to create clusters. We tested for robustness of the clusters in ERA-5 (Fig. A1) and EC-Earth (Fig. A2) by varying the number of clusters (*n_clusters* = 2, 3, 4, 5, 6, 7, and 8) and percentile thresholds (q = 80, 85, 90, and 95) used to create the binary strings. Since there are slight differences between the datasets, we also observe only small differences in the boundaries of the clustering. Because of these small differences, we decided to not use the exact same parameters as used by McKinnon et al. (2016). In the original work, the threshold was fixed at the 95th percentile, and they choose *n_clusters* = 5. For ERA-5, the exact same settings render a similar clustering result. For EC-Earth we choose the clustering output (*n_clusters* = 5, q = 50) such that the eastern U.S. cluster is most similar to the original eastern U.S. cluster found by (McKinnon et al. 2016). The final clusters are shown in Fig. 2.

APPENDIX B

Double Cross Validation

To fit and validate a statistical model, we need a sufficient amount of independent data points. Particularly for dynamics on S2S time scales, this is challenging with only 40 years of data for ERA-5. As mentioned in section 2a, we detrend all data to avoid that we are fitting a spurious signal to a long-term trend. Using the response guided approach, we make choices drawn from data, which increases the danger of overfitting (Michaelsen 1987). We can minimize this pitfall with 1) a strict train-test split throughout the whole analysis, 2) doing robustness tests such as testing different train-validation-test combinations (e.g., see Fig. F3 in appendix F). As depicted in Fig. B1, we use a stratified 10-fold cross validation to split training and test data. This means that the test years are not completely random, since the test set is forced to be a representative sample in terms of the amount of events. This helps to avoid train/test combinations that are by chance dominated by a certain phase of multiannual or decadal variability and it allows us to validate with different train/test sets, which is not possible with e.g., the leave-one-year-out method. Because we cannot reliably estimate the skill based on only 4 years of test data, we repeat the CPPA

TABLE C1. List of all SST precursor time series extracted by CPPA. The whole or a subset of the precursors are used for Figs. 5–11. They are based on the ERA-5 dataset.

ERA-5 Precursor labels/names	No.
CPPAsv	10
0..1..sst	10
0..2..sst	10
0..3..sst	10
0..4..sst	10
0..5..sst	5
0..6..sst	10
0..7..sst	7
0..8..sst	2

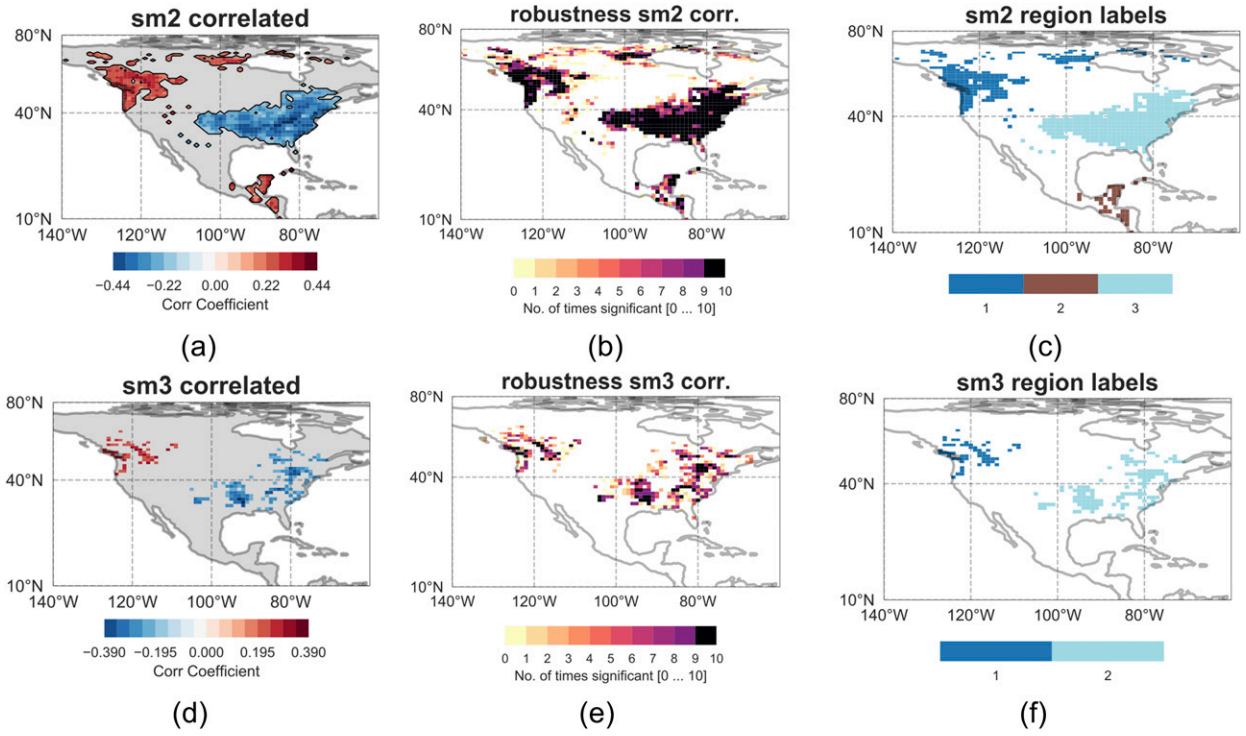

FIG. D1. Similar to Fig. C3 of appendix C, but for soil moisture.

algorithm and the subsequent model fitting 10 times. We then concatenate all forecast test years and calculate our skill metrics based on all the years in the dataset (40 years for ERA-5). Thus, we do not train a single statistical model but rather 10 slightly different ones.

APPENDIX C

CPPA versus Linear Point Correlation Map Approach

For the extracted precursors as shown in Fig. 4, we only show the mean over the training sets. However, as depicted in Fig. B1 in appendix B, we extract the precursor regions once for each training set (and for each lag), see Fig. C1. By looking at how robust the precursor region extraction was when using slightly different subsets of data, we can plot the robustness of the precursor regions (Fig. C2).

We also compare CPPA with the conventional pointwise correlation map approach. CPPA only looks at relatively extreme events (hot days) to learn the precursor regions. If the signal of the precursor only arises in the tail of the conditional temperature distribution, CPPA might enable detection of precursors showing a nonlinear relationship with eastern U.S. temperature. When comparing the output of CPPA versus the correlation map approach shown in Fig. C3, we observe a qualitatively similar pattern. This shows that either 1) the correlation map approach was still able to detect a signal when the underlying signal was in reality nonlinear, or 2) the SST relationship with temperature is by good approximation linear. We also note the correlation map shows a higher robustness compared to CPPA, which only learns from events versus nonevents. The higher robustness is also the reason to use the correlation map approach to extract soil moisture time series. Although with CPPA, we were able to stay close to the analysis as done by McKinnon et al. (2016).

Because CPPA objectively searches for precursor regions based on training data that slightly differs for each train-test split, some precursor regions are not always extracted. Table C1 shows all the precursor regions (time series) that were extracted and the count denotes how many times it is present in the 10 training sets. The format of the labels is {*lag*}..{*region label*}..{*variable name*}. The labels correspond to the labels shown in Fig. C1. Note that the lag refers to the lag at which the precursors were retrieved. Thus, we did not change the precursors as function of lag as done by McKinnon et al. (2016), since we found that using the time series of lag = 00, produced the best forecast skill. We expect this is due to

TABLE D1. As in Table C1, but for soil moisture precursors that are based on the ERA-5 dataset.

ERA-5 precursor labels/names	No.
0..1..sm2	10
0..2..sm2	10
0..3..sm2	10
0..1..sm3	10
0..2..sm3	10

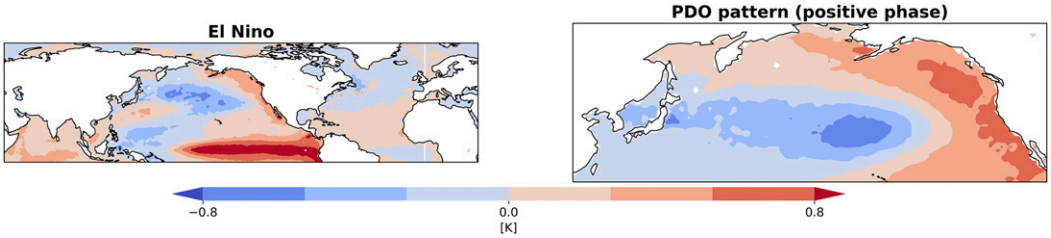

FIG. E1. (left) El Niño phase of ENSO, found by taking a composite mean where the 5-month smoothed Niño-3.4 time series exceeds 0.4°. (right) PDO pattern (mean over training sets). The retrieval is obtained by calculating the first EOF (or loading pattern) for Pacific area-weighted SST between 20° and 65°N and between 115°E and –110°W. Time series are used for the computation of the cross-correlation matrix and for the forecasts (PDO + ENSO + sm).

the fact that the signal-to-noise ratio is largest at lag = 0. The time series are subsequently shifted to match the lead time on the x axis of the verification figures.

APPENDIX D

Soil Moisture Time Series

For the final forecast we additionally add information from soil moisture layer 2 (7–28 cm) and layer 3 (28–72 cm). We choose these two deeper layers because we expect that there is more memory in the deeper layers since there is less mixing with the atmosphere. We include soil-moisture using an existing framework as introduced by Kretschmer et al. (2017) that is similar to CPPA. The soil moisture time series are retrieved by 1) calculating which grid cells are significantly correlating with the $T90_m$ time series at lag = 0, 2) subsequently clustering regions of same sign together in the same fashion as done for CPPA, and 3) calculating the area-weighted spatial mean time series for each cluster, results for this analysis are shown in Fig. D1 and Table D1.

APPENDIX E

Climate Indices

For our daily ENSO time series, we use the Niño-3.4 spatial region (5°S–5°N and 170°–120°W) to calculate the area-weighted mean of the detrended SSTA daily data (Deser and Trenberth 2016). For the calculation of the PDO time series, we first aggregate the detrended SSTA daily data to monthly means. Based on the monthly mean area-weighted SSTA training data, we construct the first EOF (or loading pattern) of the North Pacific (20°–70°N and 115°E–110°W) (Deser and Trenberth 2016). The loading pattern is projected on the (daily) test data to obtain the daily principal component time series.

We calculate the PDO with the training data for each test set (as illustrated by Fig. B1 in appendix B) to obtain an out-of-sample time series of the PDO. See Fig. E1 for the (mean over training sets) PDO pattern and a composite mean of the El Niño phase.

APPENDIX F

Supporting Information Forecasts

When we aggregate to 15-day means, without overlap in the windows, the lead time can be defined in multiple ways. To make our forecast similar to an operational implementation, the lead time is defined such that we are predicting the centered date of a time window, using only information from the past. Figure F1 shows a schematic illustration where we predict the centered date 26 August 2012. To select the precursor dates, we shift lag = 25 and the

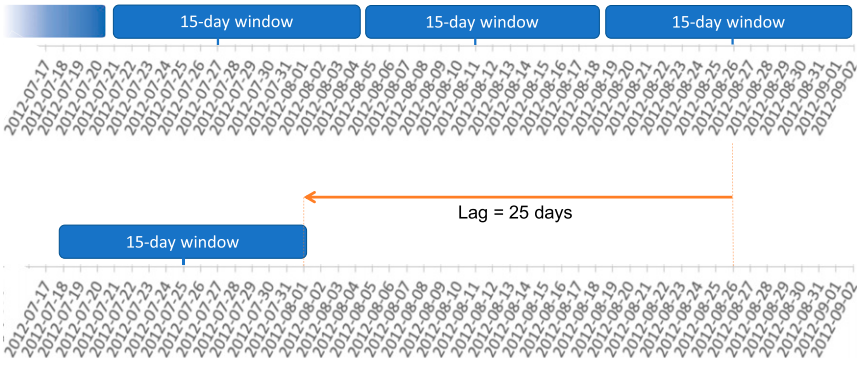

FIG. F1. Schematic illustration of the temporal aggregation and how the lead times are defined. (top) Dates represent the time series belonging to the target time series. (bottom) Dates represent the time series of the precursors.

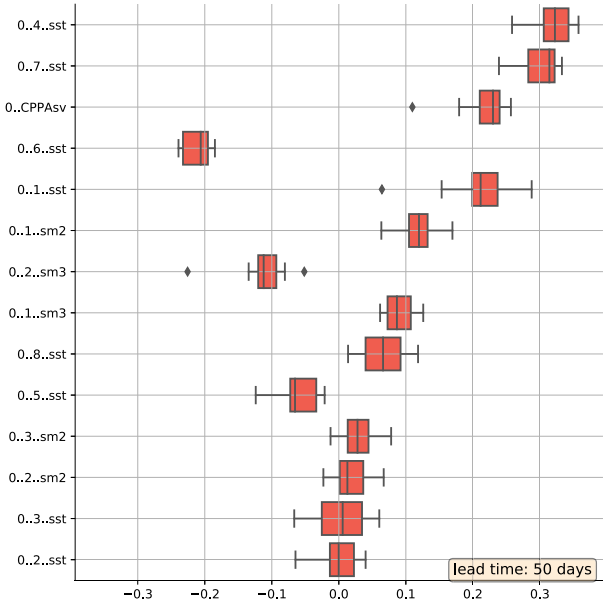

FIG. F2. Boxplot of the logistic regression coefficients that were fitted using 10 different training sets with a lead time of 50 days.

additional 15 days back in time. Hence, the prediction is made on 1 August 2012, 25 days in advance, using information of 1 August 2012 and of the previous 14 days. Note that the exact summer dates that we originally forecast on daily time scale inevitably change from 24 June to 22 August to the centered dates 27 June–26 August: exactly five bins of 15 days.

In Fig. F2, we use a boxplot to convey the consistency between models that were learned on different training datasets. The corresponding precursor regions can be found in appendices C and D. The spread in the logistic regression coefficients is generally small, indicating that overall, the models were similar. This supports that what the model learned was not a lucky fit that resulted in good skill scores on the test dataset, but rather, it relearned the same associations when applying perturbations to the training data. We will not go into discussing the physical meaning of the coefficients, since a model that provides high forecast skill, does not necessarily inform about the underlying causal structure (Li et al. 2020; Runge et al. 2019).

Figure F3 show a robustness check for the forecast skill, where we tested the influence of using 3 different combinations of train-validation sets for the "tune forecast model" step in Fig. B1 of appendix B. Sections 3c and 3d showed that, using CPPA or PEP as precursor(s), hot-day events do not show predictive skill at long leads. Figure F4 shows the forecast skill when keeping the same base rate and aggregating over time (i.e., the hot 15-day mean events). ERA-5 does not show an increase in forecast skill compared to forecasting hot-day events. EC-Earth, with many more data points, shows a small

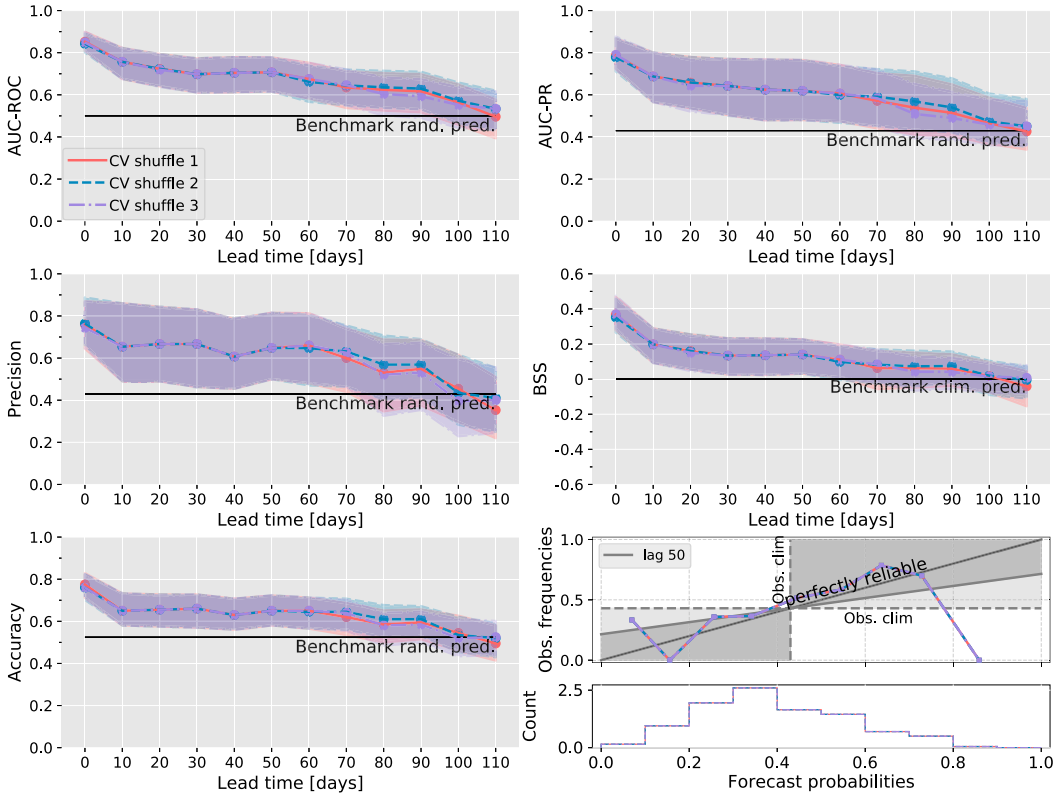

FIG. F3. Forecast-skill robustness test in which we used three different combinations of train-validation sets for the "tune forecast model" step (depicted in Fig. B1 of appendix B).

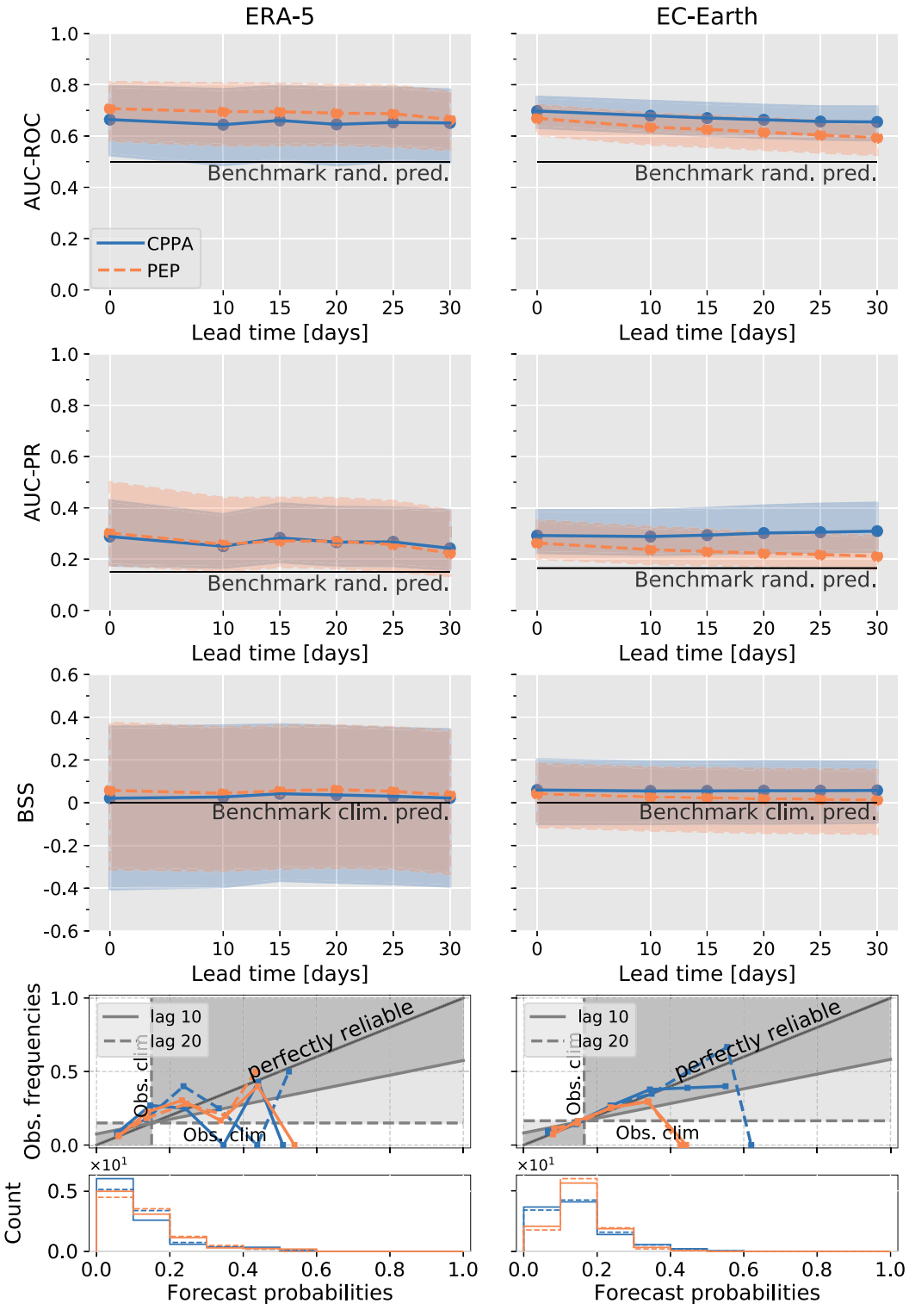

FIG. F4. Forecast validation for "hot 15-day mean events" using ERA-5 (40 years of data) and EC-Earth (160 years of data).

increase in skill relative to the daily events, but still not a significant one.

REFERENCES

Alfaro, E. J., A. Gershunov, and D. Cayan, 2006: Prediction of summer maximum and minimum temperature over the central and western United States: The roles of soil moisture and sea surface temperature. *J. Climate*, **19**, 1407–1421, https://doi.org/10.1175/JCLI3665.1.

Ardilouze, C., and Coauthors, 2017: Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability. *Climate Dyn.*, **49**, 3959–3974, https://doi.org/10.1007/s00382-017-3555-7.

Bello, G. A., M. Angus, N. Pedemane, J. K. Harlalka, F. H. M. Semazzi, V. Kumar, and N. F. Samatova, 2015: Response-guided community detection: Application to climate index discovery. *Machine Learning and Knowledge Discovery in Databases: ECML PKDD*, A. Appice et al., Eds., Lecture Notes in Computer Science, Vol. 9285, Springer, 736–751, https://doi.org/10.1007/978-3-319-23525-7_45.

Boschat, G., I. Simmonds, A. Purich, T. Cowan, and A. B. Pezza, 2016: On the use of composite analyses to form physical hypotheses: An example from heat wave–SST associations. *Sci. Rep.*, **6**, 29599, https://doi.org/10.1038/srep29599.

Cohen, J., D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, 2018: S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal (S2S) forecasts. *Wiley Interdiscip. Rev.: Climate Change*, **10**, e00567, https://doi.org/10.1002/wcc.567.

Copernicus Climate Change Service (C3S), 2017: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service, accessed 4 May 2018, https://cds.climate.copernicus.eu/cdsapp#!/home.

Deng, K., M. Ting, S. Yang, and Y. Tan, 2018: Increased frequency of summer extreme heat waves over Texas area tied to the amplification of Pacific zonal SST gradient. *J. Climate*, **31**, 5629–5647, https://doi.org/10.1175/JCLI-D-17-0554.1.

Deo, R. C., M. K. Tiwari, J. F. Adamowski, and J. M. Quilty, 2017: Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. *Stochastic Environ. Res. Risk Assess.*, **31**, 1211–1240, https://doi.org/10.1007/s00477-016-1265-z.

Deser, C., and K. Trenberth, 2016: The Climate Data Guide: Pacific Decadal Oscillation (PDO): Definition and indices. NCAR/UCAR, accessed 26 September 2019, https://climatedataguide.ucar.edu/climate-data/pacific-decadal-oscillation-pdo-definition-and-indices.

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. Rodrigues, 2013: Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, https://doi.org/10.1002/WCC.217.

Dobrynin, M., and Coauthors, 2018: Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophys. Res. Lett.*, **45**, 3605–3614, https://doi.org/10.1002/2018GL077209.

Fawcett, T., 2006: An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874, https://doi.org/10.1016/j.patrec.2005.10.010.

Finnis, J., W. W. Hsieh, H. Lin, and W. J. Merryfield, 2012: Non-linear post-processing of numerical seasonal climate forecasts. *Atmos.–Ocean*, **50**, 207–218, https://doi.org/10.1080/07055900.2012.667388.

Frankignoul, C., 1985: Sea surface temperature anomalies, planetary waves, and air-sea feedback in the middle latitudes. *Rev. Geophys.*, **23**, 357, https://doi.org/10.1029/RG023i004p00357.

Hall, R. J., A. A. Scaife, E. Hanna, J. M. Jones, and R. Erdélyi, 2017: Simple statistical probabilistic forecasts of the winter NAO. *Wea. Forecasting*, **32**, 1585–1601, https://doi.org/10.1175/WAF-D-16-0124.1.

Haustein, K., and Coauthors, 2016: Real-time extreme weather event attribution with forecast seasonal SSTs. *Environ. Res. Lett.*, **11**, 064006, http://doi.org/10.1088/1748-9326/11/6/064006.

Hazeleger, W., and Coauthors, 2012: EC-Earth V2.2: Description and validation of a new seamless earth system prediction model. *Climate Dyn.*, **39**, 2611–2629, https://doi.org/10.1007/s00382-011-1228-5.

Hessl, A. E., D. Mckenzie, and R. Schellhaas, 2004: Drought and Pacific decadal oscillation linked to fire occurrences in the inland Pacific Northwest. *Ecol. Appl.*, **14**, 425–442, https://doi.org/10.1890/03-5019.

Hewitt, H. T., and Coauthors, 2017: Will high-resolution global ocean models benefit coupled predictions on short-range to climate timescales? *Ocean Modell.*, **120**, 120–136, https://doi.org/10.1016/j.ocemod.2017.11.002.

Hodson, D. L., R. T. Sutton, C. Cassou, N. Keenlyside, Y. Okumura, and T. Zhou, 2010: Climate impacts of recent multidecadal changes in Atlantic Ocean sea surface temperature: A multimodel comparison. *Climate Dyn.*, **34**, 1041–1058, https://doi.org/10.1007/s00382-009-0571-2.

Jaccard, P., 1912: The distribution of the flora in the Alpine zone. *New Phytol.*, **11**, 37–50, https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Jaiser, R., K. Dethloff, D. Handorf, A. Rinke, and J. Cohen, 2012: Impact of sea ice cover changes on the Northern Hemisphere atmospheric winter circulation. *Tellus*, **64A**, 11595, https://doi.org/10.3402/tellusa.v64i0.11595.

Kaspi, Y., and T. Schneider, 2011: Winter cold of eastern continental boundaries induced by warm ocean waters. *Nature*, **471**, 621–624, http://doi.org/10.1038/nature09924.

Kharin, V. V., and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150, https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2.

Kornhuber, K., V. Petoukhov, S. Petri, S. Rahmstorf, and D. Coumou, 2017: Evidence for wave resonance as a key mechanism for generating high-amplitude quasi-stationary waves in boreal summer. *Climate Dyn.*, **49**, 1961–1979, https://doi.org/10.1007/s00382-016-3399-6.

Kretschmer, M., J. Runge, and D. Coumou, 2017: Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophys. Res. Lett.*, **44**, 8592–8600, https://doi.org/10.1002/2017GL074696.

Krishnamurthy, V., 2019: Predictability of weather and climate. *Earth Space Sci.*, **6**, 1043–1056, https://doi.org/10.1029/2019EA000586.

Kushnir, Y., W. A. Robinson, I. Bladé, N. M. J. Hall, S. Peng, and R. Sutton, 2002: Atmospheric GCM response to extratropical SST anomalies: Synthesis and evaluation. *J. Climate*, **15**, 2233–2256, https://doi.org/10.1175/1520-0442(2002)015<2233:AGRTES>2.0.CO;2.

Li, J., L. Liu, T. D. Le, and J. Liu, 2020: Accurate data-driven prediction does not mean high reproducibility. *Nat. Mach. Intell.*, **2**, 13–15, http://doi.org/10.1038/s42256-019-0140-2.

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the Relative Operating Characteristics (ROC) and Relative Operating Levels (ROL) curves. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, https://doi.org/10.1256/003590002320603584.

McKinnon, K. A., A. Rhines, M. P. Tingley, and P. Huybers, 2016: Long-lead predictions of eastern United States hot days from Pacific sea surface temperatures. *Nat. Geosci.*, **9**, 389–394, https://doi.org/10.1038/ngeo2687.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.*, **26**, 1589–1600, https://doi.org/10.1175/1520-0450(1987)026<1589:CVISCF>2.0.CO;2.

Miralles, D. G., A. J. Teuling, C. C. Van Heerwaarden, and J. V. G. De Arellano, 2014: Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nat. Geosci.*, **7**, 345–349, https://doi.org/10.1038/ngeo2141.

Murtagh, F., and P. Contreras, 2012: Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.*, **2**, 86–97, https://doi.org/10.1002/widm.53.

Nie, Y., Y. Zhang, G. Chen, and X.-Q. Yang, 2016: Delineating the barotropic and baroclinic mechanisms in the midlatitude eddy-driven jet response to lower-tropospheric thermal forcing. *J. Atmos. Sci.*, **73**, 429–448, https://doi.org/10.1175/JAS-D-15-0090.1.

Nobre, G. G., J. E. Hunink, B. Baruth, J. C. J. H. Aerts, and P. J. Ward, 2019: Translating large-scale climate variability into crop production forecast in Europe. *Sci. Rep.*, **9**, 1277, https://doi.org/10.1038/s41598-018-38091-4.

Putrasahan, D. A., A. J. Miller, and H. Seo, 2013: Isolating mesoscale coupled ocean-atmosphere interactions in the Kuroshio Extension region. *Dyn. Atmos. Oceans*, **63**, 60–78, http://doi.org/10.1016/j.dynatmoce.2013.04.001.

Röthlisberger, M., O. Martius, and H. Wernli, 2018: Northern Hemisphere Rossby wave initiation events on the extratropical jet—A climatological analysis. *J. Climate*, **31**, 743–760, https://doi.org/10.1175/JCLI-D-17-0346.1.

Runge, J., P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, 2019: Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.*, **5**, eaau4996, https://doi.org/10.1126/sciadv.aau4996.

Saito, T., and M. Rehmsmeier, 2015: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS One*, **10**, e0118432, https://doi.org/10.1371/journal.pone.0118432.

Schubert, E., M. Ester, X. Xu, H. P. Kriegel, and J. Sander, 2017: DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.*, **42**, 19, https://doi.org/10.1145/3068335.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: Earth-science reviews investigating soil moisture– climate interactions in a changing climate: A review. *Earth Sci. Rev.*, **99**, 125–161, http://doi.org/10.1016/j.earscirev.2010.02.004.

Seo, E., and Coauthors, 2019: Impact of soil moisture initialization on boreal summer subseasonal forecasts: Mid-latitude surface air temperature and heat wave events. *Climate Dyn.*, **52**, 1695–1709, http://doi.org/10.1007/s00382-018-4221-4.

Steptoe, H., S. E. Jones, and H. Fox, 2018: Correlations between extreme atmospheric hazards and global teleconnections: Implications for multihazard resilience. *Rev. Geophys.*, **56**, 50–78, https://doi.org/10.1002/2017RG000567.

Strazzo, S., D. C. Collins, A. Schepen, Q. J. Wang, E. Becker, and L. Jia, 2019: Application of a hybrid statistical-dynamical system to seasonal prediction of North American temperature and precipitation. *Mon. Wea. Rev.*, **147**, 607–625, https://doi.org/10.1175/MWR-D-18-0156.1.

Teng, H., G. Branstator, A. B. Tawfik, and P. Callaghan, 2019: Circumglobal response to prescribed soil moisture over North America. *J. Climate*, **32**, 4525–4545, https://doi.org/10.1175/JCLI-D-18-0823.1.

Thomson, S. I., and G. K. Vallis, 2018: Atmospheric response to SST anomalies. Part II: Background-state dependence, teleconnections, and local effects in summer. *J. Atmos. Sci.*, **75**, 4125–4138, https://doi.org/10.1175/JAS-D-17-0298.1.

Totz, S., E. Tziperman, D. Coumou, K. Pfeiffer, and J. Cohen, 2017: Winter precipitation forecast in the European and Mediterranean regions using cluster analysis. *Geophys. Res. Lett.*, **44**, 12 418–12 426, https://doi.org/10.1002/2017GL075674.

van Der Linden, E. C., R. J. Haarsma, and G. van Der Schrier, 2019: Impact of climate model resolution on soil moisture projections in central-western Europe. *Hydrol. Earth Syst. Sci.*, **23**, 191–206, https://doi.org/10.5194/hess-23-191-2019.

van der Wiel, K., N. Wanders, F. M. Selten, and M. F. Bierkens, 2019: Added value of large ensemble simulations for assessing extreme river discharge in a 2°C warmer world. *Geophys. Res. Lett.*, **46**, 2093–2102, https://doi.org/10.1029/2019GL081967.

Varoquaux, G., L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, 2015: Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Comput. Commun.*, **19**, 29–33, https://doi.org/10.1145/2786984.2786995.

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

——, and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate Atmos. Sci.*, **1**, 3, https://doi.org/10.1038/s41612-018-0013-0.

——, and Coauthors, 2019: Sub-seasonal to seasonal prediction of weather extremes. *Sub-Seasonal to Seasonal Prediction*, A. W. Robertson and F. Vitart, Eds., Elsevier, 365–386, https://doi.org/10.1016/b978-0-12-811714-9.00017-6.

Wang, H., S. Schubert, M. Suarez, and R. Koster, 2010: The physical mechanisms by which the leading patterns of SST variability impact U.S. precipitation. *J. Climate*, **23**, 1815–1836, https://doi.org/10.1175/2009JCLI3188.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wirth, V., M. Riemer, E. K. M. Chang, and O. Martius, 2018: Rossby wave packets on the midlatitude waveguide—A review. *Mon. Wea. Rev.*, **146**, 1965–2001, https://doi.org/10.1175/MWR-D-16-0483.1.

WMO, 2006: Standardized Verification System (SVS) for long range forecasts. WMO Doc., 17 pp., https://www.wmo.int/pages/prog/www/DPS/LRF/ATTACHII-8SVSfrom%20WMO_485_Vol_I.pdf.

——, 2017: Annual report: Climate risk and early warning systems. WMO Climate Risk and Early Warning Systems Tech. Rep., 23 pp.